International Journal of Interactive Multimedia and Artificial Intelligence

December 2021, Vol. VII, Number 2
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

“Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution.”

Albert Einstein

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 7, NUMBER 2

EDITORIAL TEAM

OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/Science Edition, Current Contents®/Engineering Computing and Technology.*

COPYRIGHT NOTICE

Editor’s Note

The International Journal of Interactive Multimedia and Artificial Intelligence - IJIMAI - provides a space in which scientists and professionals can report about new advances in Artificial Intelligence (AI). On this occasion, for the last edition of the year, I am pleased to present a regular issue including different investigations covering aspects and problems in AI and its use in various fields such as medicine, education, image analysis, protection of data, among others.

On behalf of the editorial board, I must once again thank the reviewers and authors who trust and support the journal with their work.

I am going to briefly introduce each of the works that we will find in this edition.

The volume begins with a paper that focuses on mental health, Daus and Backenstrass carry out a pilot study with bipolar patients that assesses the viability of the new care approach through the recognition of emotions using mobile phones. The results indicate that it could be easy to analyze verbal emotions and facial recognition without compromising the privacy of the patient.

The next two articles are also focused on the health sector.

Amador-Domínguez et al. propose a hybrid system between Case Based Reasoning (CBR) and Deep Learning (DL) for the generation of medical reports. The authors take advantage of the CBR's explicability and the DL predictive power. The proposed system is fully modular and adaptable to various clinical scenarios, specifically the authors present a case focused on the development of radiology reports to illustrate the proposal.

On the other hand, Bareño-Castellanos et al. develop a prototype that supports medical decisions with patients with cardiovascular risk. Using easily obtained variables such as pressure force, weight and body mass index, different algorithms such as k-means, c-means and support vector machine (SVM) were tested with very good results, especially when combining these two last algorithms.

We continue with two works that refer to predicting human activities and gestures.

Li et al. implement a model based on deep neural networks to classify 4 affective stimuli: nervous, calm, happy and sad, using eye-tracking signals. Eye tracking has been applied in recent years in neuromarketing, neurocognition and user experience, among others. The results show that the accuracy of the model reaches 87.2%. In the future, affective computing will be applied in a wide range of fields such as medical rehabilitation, assistance for people with disabilities and education.

Verma and Singh investigate different architectures that allow to predict activities such as walk, sit down, standup, pickup, carry, throw, push, pull, wave hands and dap hands from images with depth and the sequence of skeletal joints. The proposed work is done in three different levels. In the first level, spatial-temporal features are extracted from different modalities, in the second level three independent SVMs are trained using the features extracted from the first level, and in the last level the probability scores are fused and optimized using two evolutionary algorithms such as a genetic algorithm (GA) and particle swarm optimization (PSO). They achieve 96.50% accuracy results.

Kim, Oh and Heo compare the two VGG architectures ResNet and SqueezeNet, with different convolutional filters, in order to identify different mosquito types using audio data and the circadian rhythm of insects. Adding activity circadian rhythm information to the networks showed an average performance improvement of 5.5%. The VGG13 network with 1D-ConvFilter achieved the highest accuracy of 85.7%.

In the next article, Suruliandi, Kasthuri, and Raja present a novel method named the Similarity Matrix-based Noise Label Refinement (SMNLR), which effectively predicts the accurate label from the noisy labeled facial images. In addition, they use Convolutional Neural Networks (CNN) for feature representation. The convolution and pooling layers of the CNN are able to get enough information, such as the edges, orientations, and corner features of facial images. Extensive experiments are carried out with three databases, surpassing the results of previous works.

Hernandez-Olivan et al. propose a general method to preprocess musical piece inputs for later identification. Preprocessing allows you to establish the most efficient combination of inputs to a CNN. They use a max-pooling of factor 6 at the beginning of the process as a pooling strategy, thus generating the self-similarity matrices that enter the CNN. This method reduces preprocessing and training time for the neural network.

In order to evaluate the quality of the software automatically, Gupta and Chug take into account the ISO / IEC 25010: 2011 standard where 8 characteristics of product quality are described, and maintainability is one of the most important. Maintainability refers to how easy it is for the software to deal with new requirements. To meet the objective, five Boosting Algorithms (BA) are compared using 7 empirically collected open source data sets. Based on the residual errors obtained, Gradient Boosting Machine (GBM) is the best performer, followed by Light Gradient Boosting Machine (LightGBM) for Root Mean Square Error (RMSE), whereas, in the case of Mean Magnitude of Relative Error (MMRE), eXtreme Gradient Boosting (XGB) performed the best for six out of the seven datasets.

There are also two works focused on speech. Let us see what they are about.

Nosek et al. work on synthesizing speech in multiple languages, in multi-speaker voices and multiple styles to train a neural network that identifies similarities and differences between speakers and establish relationships between the phonemes of different languages and produce high-quality synthetic speech. Using vocoders, it was shown to be capable of producing good quality synthetic speech even in languages in which it was not trained.

Automatic audio-visual speech recognition is an emerging research field where visual speech is recognized through face detection, Region of Interest (ROI) detection, and lip tracking. Debnath and Roy propose a new method for the extraction of visual characteristics using Pseudo Zernike Moment (PZM) to follow the movement of the lips. Subsequently, they weigh the importance of the characteristics with statistical analysis of Analysis of Variance (ANOVA), the Kruskal-Wallis test and the Friedman test, demonstrating that this step is important to overcome the results of other speech recognition methods. Finally, they recognize audio-visual speech using machine learning algorithms such as SVM, artificial neural networks (ANN) and Naive Bayes, with very good results.

Basavaraju et al. present a Neighborhood Structure-Based Model that locates the region where there is text in both images and videos. The tests are performed on 5 different image datasets and video frames where there is low contrast, composite background and lighting effects. In the tests we can see how the probable textual spaces are successfully localized in multiple languages.

Seal et al. propose a non-Euclidean similarity measure that is based on Jeffreys nonlinear divergence (JS). They analyze the method with

DOI: 10.9781/ijimai.2020.11.010

real and synthetic databases, demonstrating the superiority of the method over other c-means algorithms. It could be very useful for designing new clustering algorithms.

The next article is a contribution to the area of copyright protection, so important in our times. Kumari et al. optimize Discrete Wavelet Transform (DWT) for embedding an imperceptible and a robust non-blind image watermarking. The extraction is processed by a Recurrent Neural Network based Long Short-Term Memory (RNN-LSTM), obtaining the original image.

Kadry et al. recommend a methodology to solve the multi-thresholding problem of RGB scale images using entropy value. The authors propose a random search along with a novel Multiple-Objective-Function (MOF), to maximize MOF. The tests confirm that the performance of MFO is better than PSO, BA and FA, and approximately similar to MA and AOA.

Abdulkareem et al. conduct a review and analyze the most relevant studies in the image dehazing field. They conducted an objective image quality assessment experimental comparison of various image dehazing algorithms and reflect different observations that can serve as a useful guideline for practitioners who are looking for a comprehensive view on image dehazing.

Garrido et al. analyze different multi-agent systems, which implement creative methods such as establishing analogies with known problems, brainstorming, lateral and parallel thinking. The authors propose different guides for each of these methods that help determine if a solution meets the requirements of the creative task. In addition, they define a conceptual model to implement a creative computational system.

The following work is an advance towards the metaverse that we have heard so much today. Lopez et al. have thought of democratizing virtual reality experiences for the educational field. They use the concept of mixed reality that combines virtual reality and the real world in the same scenario. The proposed solution enables content creators to design, build and publish training experiences in the cloud, using Microsoft HoloLens2. The user can see an object in 3D in the same point of view as in the real environment and interact with the object thanks to the recognition of gestures and speech of HoloLens2.

Fei Xu et al. measure the similarity between works of art or paintings, extracting characteristics using the Sparse Metric Learning-based Kernel Regression (KR-SML) algorithm. Although its main objective is to improve the teaching-learning in the identification of works of art, the work also seeks to predict the genre, the artist and the style of painting. The model obtains good prediction results by combining SVM with ANN, and Human-Computer Interaction (HCI).

Cobos-Guzman et al. present a virtual assistant, capable of interacting with the public to improve communication. The system can recognize the level of attention from audiovisual resources and synchronizes the assistant to increase the level of attention of the audience. The system is composed of two large modules, one that captures audio and images to extract characteristics that represent behavior and attitudes; and another that represents knowledge through ontologies, and with the help of reinforcement learning, the assistant is able to decide the best strategy.

The following work arises as a concern resulting from the online tests used in colleges and universities, forced by the Covid-19 pandemic. Balderas et al. propose a model to detect fraudulent collaborations between students who present online assessments, based on the time of presentation of the test and the grade obtained. The software tool proved to be very useful for teachers to detect fraud.

Tlili et al. applied Learning Analytics (LA), which focuses on understanding students' in-game behavior trajectories and personal learning needs during the game. This systematic literature review examined how LA in educational games has evolved. The results indicate that factors such as student modeling, iterative design, and customization must be taken into account. Furthermore, the use of LA creates several technical challenges such as data management and ethics that are still unsolved.

And finally, in the last work Amo et al. raise a framework with 7 principles: legality, transparency, data control, anonymous transactions, responsibility, interoperability and local first. This framework is proposed to be considered in the development and adoption of educational technology that collects, stores, manages and analyzes educational data; improving the privacy and security of the data.

Dr. Xiomara Patricia Blanco Valencia
Managing Editor
Universidad Internacional de La Rioja

TABLE OF CONTENTS

Feasibility and Acceptability of a Mobile-Based Emotion Recognition Approach for Bipolar Disorder

H. Daus[1,2,3], M. Backenstrass[2,4] *

[1] Psychotherapy Practice, Dipl.-Psych. Henning Daus, Ellhofen (Germany)
[2] Institute of Clinical Psychology, Centre for Mental Health, Klinikum Stuttgart (Germany)
[3] Faculty of Science, Eberhard Karls University Tübingen (Germany)
[4] Department of Clinical Psychology and Psychotherapy, Institute of Psychology, Ruprecht-Karls-University Heidelberg (Germany)

Received 26 October 2020 | Accepted 9 June 2021 | Published 30 August 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

Over the past years, the mobile Health approach has motivated research projects to develop mood monitoring systems for bipolar disorder. Whereas mobile-based approaches have examined self-assessment or sensor data, so far, potentially important emotional aspects of this disease have been neglected. Thus, we developed an emotion-sensitive system that analyzes the verbal and facial expressions of bipolar patients in regard to their emotional cues. In this article, preliminary findings of a pilot study with five bipolar patients with respect to the acceptability and feasibility of the new approach are presented and discussed. There were individual differences in the usage frequency of the participants, and improvements regarding its handling were suggested. From the technical point of view, the video analysis was less dependable than the audio analysis and recognized almost exclusively the facial expressions of happiness. However, the system was feasible and well-accepted. The results indicate that further developments could facilitate the long-term analysis of expressed emotions in bipolar or other disorders without invading the privacy of patients.

Keywords

Assistance Systems, Bipolar Disorder, Emotion Recognition, Intelligent Systems, Mobile Health.

DOI: 10.9781/ijimai.2021.08.015

I. Introduction

Emotional expressions are an important indication of how we are feeling. Thus, sudden or severe changes in those expressions can be signs for psychopathological syndromes or diseases. One of the most common mental disorders worldwide is bipolar disorder [1], which is defined by its depressive or (hypo-) manic episodes [2]. These episodes are characterized by typical changes in the emotional experiences of patients: For example, patterns of elevated happiness and anger during mania or sadness and disgust during depression [3]. Beyond that, patients with bipolar disorder often show emotion-related difficulties like deficits in emotion regulation [4]–[8] and recognition [9]–[12] or in the afore-mentioned expression of emotions [13], [14]. Therefore, in bipolar disorder, the processing of emotional information and the verbal or facial expressions following emotional stimuli might be influenced by mood-dependent biases or regulation strategies like emotional avoidance [15]–[17]. Since all of these areas are important to the everyday lives of patients, the deficits affect their overall functioning [18]–[20].

Depending on the severity and the course of this disease, its personal and economic consequences can be enormous [21], [22]. Thus, within patient care, strategies that complement the medical and psychological treatment are of great interest [23], [24]. Over the past years, the increasing technical possibilities have strengthened the *mobile Health* (mHealth) approach in general and, just as much, in bipolar disorder [25], [26]. Especially the advantages regarding the availability, accessibility and cost efficiency of technical systems show their potential to benefit the ambulatory care system. Many mobile-based approaches for bipolar disorder include self-assessment or sensor data in mood monitoring or recognition approaches [27]–[33]. Yet, none of the existing systems have focused on the emotional aspects of this disease [34]. The expressed emotions of patients, however, could be an important indication of their emotional reactivity [14] and possibly of their affective state or recurring episodes [34].

Within the research project *Emotion-sensitive Assistance systems for the reactive psychological Interaction with people* (EmAsIn), an emotion-sensitive mHealth approach for bipolar disorder has been developed. The system, which has already been described in detail [34], combines the common self-assessment (SA) or sensor approaches with other, innovative features. It incorporates an emotion-sensitive module that retrieves audio and video data from short and actively user-triggered recordings of patients. The auditive and visual information is analyzed in respect to emotional cues or content [34]. Beyond that, the system includes optional, external assessments or third-party assessments (TPA), which are realized by involving close contacts into the ambulatory approach.

In this article, we examine the feasibility and acceptability of several system features. Therefore, data from a pilot study with patients with

* Corresponding author.

E-mail addresses: mail@praxis-daus.de (H. Daus), m.backenstrass@klinikum-stuttgart.de (M. Backenstrass).

bipolar disorder is presented and discussed. As indications regarding the feasibility of the emotion-sensitive module, the SA or the TPA approach, some usage-related aspects (like the completion of the participation periods or the frequency of feature usage) were assessed. Concerning the emotion-sensitive module, the accuracy of the word and face recognition, the detected emotional expressions and the convergence between the emotion recognition within both sources (auditive and visual) were seen as further indications of its feasibility. In order to examine the acceptability of the features, interviews with the participants were conducted. All features were expected to be feasible and, in accordance with research suggesting a higher technical affinity of younger patients [36], [37], a higher feasibility with younger participants was expected (A). In more detail, a good usability (A.1) and convergent results of the recognition approach (A.2) were expected. Since participants of an earlier interview study of the authors had been open towards the innovative features [35], all features were expected to be well-accepted (B).

II. Methods

A. Participants

From April to December 2018 five ambulatory patients with remitted bipolar disorder participated in the pilot study for an individual duration of several weeks (see results section). The study sample included three men and two women. Four participants, according to DSM-V, could be diagnostically categorized into bipolar I disorder, the fifth person into bipolar II disorder. The participants (between 24 and 51 years old) had an average age of 39.40 years (*SD* = 9.94).

B. System Features

The most important study instruments were smartphones with a pre-installed system application and a wearable device that participants wore on their wrists. The overall system acquires data from multiple sources in order to recognize mood state changes of patients. Next to the features described below, the system analyzes physiological or sleep data (e.g., heart rate or sleep duration) and behavioral patterns (e.g., recognized activities or smartphone usage). Since the overall concept of the system has been presented on another occasion [34], the following sections were reduced to essential technical aspects of the emotion recognition approach and the other two examined features.

The application analyzes intentionally initiated recordings of its users by activating the smartphone camera and microphone, as soon as this function is selected. Users then are instructed to tell their *Story of the Day* (SotD), a narrative of an important event of their day. In the study setting, smartphone holders and external microphones were included to ensure a good quality of the recordings. Only if the user approves, the system saves the recording locally and analyzes the given information. In that case, automatic transcriptions of the verbal information are analyzed in regard to the word count of different emotional categories by following the *Linguistic Inquiry and Word Count* (LIWC) approach [38]. For example, the specific categories of sadness, anxiety or anger include words like “to regret”, “nervous” or “outraged”. The more comprehensive categories of positive or negative emotions contain words of all the specific, emotional word categories with positive or negative connotations. The visual information, on the other hand, is automatically separated into multiple video frames per second. All captured facial expressions then are analyzed regarding the presence of one of four basic emotions (happiness, sadness, anger and anxiety) by following the *Facial Action Coding System* (FACS) [39].

The app-based SA approach has also been developed within the research project and has not been clinically evaluated yet. It includes six 7-point items regarding the most important symptoms of bipolar disorder. All items can be assessed by choosing values from -3 to 3 with negative values or positive values representing depressive or (hypo) manic symptoms. The related and newly developed TPA of partners or related parties are realized using separate web applications (and are, therefore, not necessarily assessed via smartphone). They also include six items that are almost identical with the SA items, and are answered using the same scale [34].

The participants of the pilot study were asked to keep their daily routines as they were and to integrate the assistance system as far as possible. Whereas some system components were designed to be optional [34], the three presented features (or, two features in case of missing third parties), should have been used regularly (e.g., on a daily basis). The SotD was supposed to take about two minutes per day and to contain narrations of recent events that had affected the emotional experiences of the narrator. Furthermore, the stories were supposed to be recorded in a well exposed room without the presence of other persons and with a frontally placed camera (smartphone). During the pilot study, the assistance system was not yet fully developed and did not automatically react to the assessed data or recordings.

C. Semi-structured Interview

Potential technical issues and the individual usage experiences were examined at the end of the study participation during semi-structured interviews. The interview guide included questions about all system features. The section about the emotion recognition approach was especially detailed and concerned aspects like the perceived emotional intensity or potential confounding factors while telling the SotD. For example, the participants were asked, “Do you think that your perceived emotions or your emotional expressions were influenced by the missing dialog partner?” All questions could be answered on 5-point scales from 1 “negative” or “not at all” to 5 “positive” or “exactly”. Beyond that, every issue allowed for explanations or comments. The nine interview questions that are relevant to this article (seven of them SotD-related) took about ten minutes.

D. Indicators of Feasibility and Acceptability

The measurement of the feasibility was operationalized through several usage-related aspects: The duration of the study participation of all patients (in reference to the targeted duration, see section E) was seen as an indicator regarding the feasibility of the overall system. Furthermore, for all features, the usage frequency (in reference to the study duration) was assessed and seen as a feasibility measure. As for the SotD module, the duration of the recordings was seen as an additional measure of their usability. Moreover, the technical functionality of this module was assessed in order to infer its (technical) feasibility. Thus, the count of recognized words was assessed, and the congruence of automatic transcriptions with a sample of manual transcriptions was analyzed (see section E). The count of the recognized faces was assessed and compared to the count of video frames in order to evaluate the recognition. The amount of the recognized emotional expressions of several emotional categories (in reference to the count of words and frames with faces) was assessed and seen as a further (technical) feasibility measure of the module. Finally, by analyzing the congruence of the emotion recognition results of both sources, the measure of their convergent validity was included as a feasibility indicator, because it could be seen as a potential proof of concept. The acceptability of the features was operationalized and assessed through questions of the interviews. Therefore, the selection of the participants regarding the possible answer choices or additional comments delivered the acceptability measures.

E. Design and Analysis

As for most of the research questions, the five participants were considered single cases and their data was analyzed separately.

However, in spite of the small number of participants, in regard to some specific questions (e.g., the age effect or the acceptability), correlation analyses or other methods were applied within a statistical group design. The long-term assessment of the feasibility data (see section D) was realized within the ambulatory environment of the patients (ambulatory assessment). An individual assessment period of twelve weeks (about 84 days) per participant was aimed for. With respect to the potential age effect on the usage behavior, the correlations between the age of the participants at the beginning of the study and the frequency of their SA or the SotD assessments were calculated. The congruence between the automatic SotD transcriptions and a sample of ten manual transcriptions (as a feasibility measure of the emotion-sensitive approach) was assessed by comparing all recognized LIWC [38] word categories (e.g. pronouns, numbers or specific themes like "money") within both sources. The percentage of the recognized emotional expressions within all recognized words and faces of each SotD assessment was analyzed following the LIWC [38] and FACS [39] approaches. The congruence of both recognition sources (convergent validity) was analyzed by calculating the correlations between corresponding or contradicting LIWC and FACS categories. The interviews, which were developed to assess acceptability aspects, were conducted in German with a single measurement time point at the end of the study participation. The answers as well as the comments were recorded, translated into English and delivered quantitative and occasional qualitative data. All descriptive or correlative analyses were conducted using *Microsoft Excel V. 16.40* and *IBM SPSS Statistics V. 26*.

III. Results

A. Feasibility

The individual study duration of the five participants was at least 57 and up to 134 days. On average, the patients of the pilot study participated for 87.40 days (*SD* = 32.14). As illustrated in Table I, the individual study duration of some participants considerably exceeded or fell below the targeted assessment period.

TABLE I. Study Duration, Assessment Numbers and SOTD[a] Duration for Each Participant

Code	Study duration in days	SA[b]	TPA[c]	SotD	
		N	N	N	Duration in s (M, SD)	
P1	134	110	1	14	94.36	38.73
P2	97	100	0	43	92.81	36.42
P3	92	46	17	15	104.47	31.91
P4	57	64	0	6	96.17	47.91
P5	57	54	0	48	148.02	43.91

[a]SOTD = Story of the Day; [b]SA = self-assessments; [c]TPA = third-party assessments.

1. Usage-related Aspects

The individual assessment numbers for the SA, TPA, and SotD for each participant are also shown in Table I. On several occasions, the same features were used several times during the same days. That is why, for example, one participant completed 64 SA within 57 days (see Table I). According to the study instruction of daily usage, Fig. 1 illustrates solely the percentage of participation days for each patient, on which they (or their related parties) used the three concerned features at least once. In reference to the SA, the usage frequency for all participants ranged from 40.22 % to 96.49 % of the participation days. The TPA were only used by related parties of two of the participants, resulting in a minimum of 0 % usage and a maximum of 17.39 % of the participation days. For the SotD assessments, the usage frequency ranged from 8.21 % to 82.46 %. Together, the five participants of the pilot study recorded 126 SotDs. As one can see in Table I, the individual mean values for the audio file duration ranged from 92.81 s to 148.02 s.

There were strong, negative correlations between the age of the participants and the percentage of days, on which they used the SA [*r*(3) = -.60, *p* = .285] and the SotD [*r*(3) = -.70, *p* = .188].

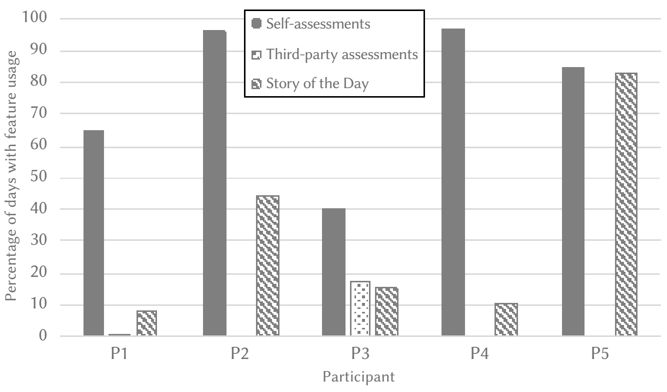


Fig. 1. Usage frequency for specific system features and each participant in reference to the individual study duration (in days).

2. Functionality of the Emotion Recognition Approach

Within the automatic transcriptions of the SotD audio files of each participant, on average, the LIWC program [38] recognized between 106.29 and 240.77 words (see Table II). A random sample of ten audio files was also transcribed manually and then analyzed with the LIWC software. All recognized LIWC word categories (not only emotional categories) within the manual and the corresponding automatic transcriptions matched from 72.00 % to 93.00 % (*M* = 87.30, *SD* = 6.70). The LIWC analysis of the audio data of all participants found, on average, between 3.93 and 11.15 single words corresponding to the category of positive emotions. This means that, for example, all audio files of one participant contained on average 5.47 single words belonging to that specific category. The corresponding mean values for each participant are illustrated in Table II. For the category of negative emotions, the individual mean values ranged from 1.33 to 4.52 words per file (see Table II). Regarding the more specific emotional category of sadness, the LIWC analysis resulted in mean values between 0.43 and 1.04 words per file (Table II). For the other two afore-mentioned, specific emotional LIWC categories (anxiety and anger), all corresponding mean values resulting from the analysis were ≤ 0.67.

TABLE II. LIWC[a] Analysis of the SOTD[b] Recordings for Each Participant

Code	Number of words		Positive Emotions		Negative Emotions		Sadness	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
P1	106.29	81.40	3.93	3.17	2.21	1.67	0.43	0.51
P2	140.02	59.40	5.93	2.87	1.91	1.82	0.53	0.74
P3	173.33	58.17	5.47	3.20	3.00	2.56	0.93	1.16
P4	149.67	89.91	6.00	3.85	1.33	1.21	0.83	0.98
P5	240.77	92.79	11.15	4.87	4.52	2.42	1.04	1.01

[a]LIWC = Linguistic Inquiry and Word Count; [b]SotD = Story of the Day; P1 (*N* = 14), P2 (*N* = 43), P3 (*N* = 15), P4 (*N* = 6), P5 (*N* = 48).

In percent (compared to the total wordcount of the audio files and for each participant separately), there were average amounts from 3.00 % to 5.89 % of words belonging to the category of positive emotions in the recordings (see Fig. 2). In regard to the category of negative emotions, the individual mean values ranged from 0.93 % to 2.01 % of words per file. The more specific category of sadness accounted for 0.34 % to 0.58 % of the wordcounts (see Fig. 2).

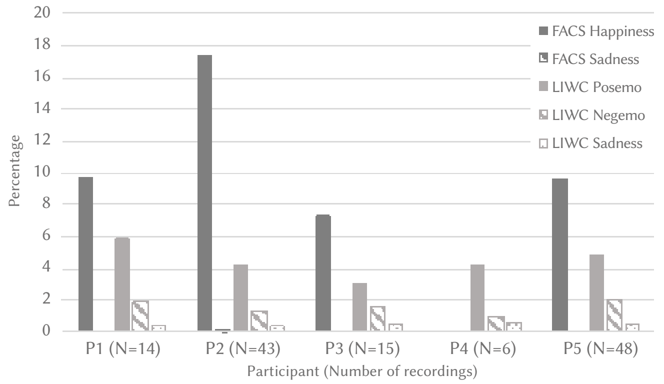


Fig. 2. Percentage of recognized emotions of Facial Action Coding System (FACS) categories and Linguistic Inquiry and Word Count (LIWC) categories in the Story of the Day recordings in reference to the individual number of frames with recognized faces (FACS) or the individual word count (LIWC).

The corresponding SotD video files were separated by the assistance system into, on average and for each participant separately, 2787.42 to 4443.42 video frames. For two assessments, the smartphone application apparently did not work accurately, and no video data was recorded. Both assessments belonged to P1 and had to be excluded from the calculations concerning the FACS [39] analysis. With four participants, within more than 94 % of the video frames there were recognized faces. As for P4, only 0.31 % of the video frames contained recognized faces. Thus, the individual mean values of frames with recognized faces that could be analyzed by the FACS-based software ranged from 7.00 to 4438.25 (see Table III). When the 6 video files of P4 without face recognition are not taken into account, for the remaining four participants, the FACS analysis found individual mean values of 209.93 to 538.51 frames with recognized happiness in the facial expressions (see Table III). However, the automatic FACS analysis detected almost no sadness in the facial expressions, resulting in individual mean values of all participants of ≤ 0.60 frames with recognized sadness (see also Table III). Beyond that, the FACS analysis detected none of the other two examined emotional FACS categories (anxiety or anger). Regarding the participants with successful face recognition, the individual mean percentage of frames with happiness out of all relevant frames (with faces) ranged from 7.38 % to 17.44 % (see Fig. 2). The mean percentage of sadness for all participants was ≤ 0.02 % (see Fig. 2).

TABLE III. FACS[a] Analysis of the SOTD[b] Recordings for Each Participant

Code	Frames with faces		Happiness		Sadness	
	M	*SD*	*M*	*SD*	*M*	*SD*
P1	2919.17	1051.87	243.42	205.14	0.00	0.00
P2	2749.60	1122.88	538.51	638.62	0.60	3.81
P3	2961.33	963.49	209.93	219.27	0.00	0.00
P4	7.00	12.25	0.17	0.41	0.00	0.00
P5	4438.25	1315.08	435.88	384.13	0.02	0.14

[a] FACS = Facial Action Coding System; [b] SOTD = Story of the Day; P1 (*N* = 12), P2 (*N* = 43), P3 (*N* = 15), P4 (*N* = 6), P5 (*N* = 48).

In order to examine, if the measures for the verbal and facial emotional expressions had delivered convergent results, the correlations between the most important LIWC and FACS categories were analyzed for each participant separately. For all participants with working face recognition, there were moderate or strong correlations between specific verbal and facial emotion recognition categories (see Table IV for all correlations). In regard to P1, the most prominent results were strong negative and significant correlations between the FACS category happiness and the LIWC category positive emotions (see Table IV). Thus, in this case, a higher amount of recognized expressions of happiness within the video frames coincided with less emotionally positive words within the LIWC analysis.

Concerning P2, the most prominent results were strong positive and significant correlations between FACS happiness and LIWC positive emotions or between FACS sadness and the LIWC sadness categories (see Table IV). Consequently, more video frames with expressed happiness coincided with a higher amount of positive words. And more recognized sadness within the FACS analysis was associated with more sadness-related words in the LIWC analysis. As for P3, there were no significant results, and there was only one moderate correlation indicating that more facially expressed happiness was associated with the use of more emotionally positive words (see Table IV). Regarding P5, there were several moderate and significant correlations: The FACS happiness categories were positively associated with LIWC positive emotions and negatively associated with LIWC negative emotions (%). Therefore, more recognized happiness within the facial expressions coincided with more emotionally positive and less emotionally negative words within the spoken language of this participant.

B. Acceptability

During the semi-structured interviews at the end of their individual study participation, the five participants showed positive attitudes towards the app-based SA (*M* = 3.80, *SD* = 1.10). One participant, nevertheless, emphasized the relevance of external assessments due to a potentially less-reliable self-perception (see Table V). Yet, only two participants were able to use the TPA with the help of related parties. These two participants perceived the assessments very positively (*M* = 4.50, *SD* = 0.71). Other participants still commented on this feature and expressed their interest in external assessments that could compensate for biased retrospective assessments regarding their mood (see Table V). With respect to the SotD module, the overall view of the five participants was relatively positive (*M* = 3.40, *SD* = 1.14) and comments mentioned positive “side-effects” of this feature (see Table V). The instruction of this module was easy to understand (*M* = 4.60, *SD* = 0.55), and most participants could imagine using this feature on a regular basis (*M* = 4.00, *SD* = 0.71). However, comments differed regarding the potential frequency of this usage or included recommendations for possible modifications (see Table V).

When asked, if their perceived emotions during the SotD recordings and during the described current events were of a similar intensity, the participants showed a moderate agreement (*M* = 2.80, *SD* = 1.30). They further agreed that the missing dialog partner, to some degree, had influenced their perceived and expressed emotions during the SotD (*M* = 3.40, *SD* = 1.14). Comments specified that some participants had experienced the strange feeling of being observed while talking into the camera. Yet, they further mentioned that this effect was not persistent (see Table V for comments). The “selfie mode” during the recordings was perceived as somewhat influencing, too, but less strong than the missing interaction (*M* = 3.00, *SD* = 1.23). However, some participants were somewhat irritated by seeing their own emotional expressions (see Table V). When asked, if they felt burdened by the SotD recordings, the participants did not agree (*M* = 2.00, *SD* = 1.41).

TABLE IV. Pearson Correlations Between FACS[a] and LIWC[b] Categories

Codec	FACS[a]	LIWC[b]					
		Positive Emotions	Positive Emotions (%)	Negative Emotions	Negative Emotions (%)	Sadness	Sadness (%)
P1	Happiness	**-.73****	-.36	-.29	.06	-.03	.30
	Happiness (%)	**-.72****	-.12	-.36	.02	-.09	.20
P2	Happiness	**.49****	.14	.25	.02	.07	-.06
	Happiness (%)	.28	.13	.15	.04	.04	-.03
	Sadness	.00	-.03	.18	.14	**.53****	**.35***
	Sadness (%)	.00	-.03	.19	.14	**.53****	**.36***
P3	Happiness	.31	.18	.21	.05	.19	.10
	Happiness (%)	.14	.15	.00	-.09	-.07	-.12
P5	Happiness	**.44****	.05	.05	**-.31***	-.04	-.19
	Happiness (%)	**.37****	.18	-.08	**-.31***	-.08	-.16
	Sadness	.09	.17	-.21	-.20	-.15	-.15
	Sadness (%)	.09	.17	-.21	-.20	-.15	-.15

[a]FACS = Facial Action Coding System; [b]LIWC = Linguistic Inquiry and Word Count; [c]P4 excluded due to failed face recognition, for P1 and P3 the FACS analysis did not detect any sadness; P1 (*N* = 12), P2 (*N* = 43), P3 (*N* = 15), P5 (*N* = 48); ** Correlation is significant at the 0.01 level (2-tailed); * Correlation is significant at the 0.05 level (2-tailed).

TABLE V. Comments of the Participants on Selected Issues

Interview section and issue		Comment and participant number
Self-assessments		*“It´s a good feature, but I didn´t use it regularly. My self-perception isn´t that good, thus, third-party assessments are very important.”* P1
Third-party assessments		*“I would recommend using that feature and keeping it simple. Maybe one could also notify the partner to facilitate simultaneous comparisons.* P1 *“If you would ask me how I felt last week, I wouldn´t remember. Thus, third-party assessments would be helpful [in regard to future systems].”* P5
Story of the Day	**Overall view**	*“It helps your self-perception and forces yourself to observe yourself.”* P3
	Regular usage	*“If it was easier to use.”* P1 *“Two or three times a week, yes. Seven times would be difficult. It´s mentally challenging and you don´t have a story to tell each day.”* P3 *“In theory, yes. I don´t like filming myself, but if there was a feedback and it would help my situation, yes.”* P4 *“If there was a feedback, yes. But not on a daily basis.”* P5
	Missing dialog partner	*“It irritates, yes. I feel strange when I´m being filmed, it would be different with a real counterpart.”* P4 *“At the beginning it´s strange to talk to the camera, because you feel observed. Later it´s like keeping a diary.“* P5
	Selfie mode	*“I tried to focus on the camera, because it irritated me to see my own feelings on screen.”* P5

IV. Discussion

The mHealth approach offers many opportunities for the ambulatory care system and the disease management of patients with bipolar disorder. Within the EmAsIn project, to our knowledge, the first emotion-sensitive assistance system for bipolar disorder was developed and examined. The following sections of the current article first address the presented results regarding the (technical) feasibility and acceptability of the newly developed system. Subsequently, the implications of these findings with respect to clinical practice and further research as well as methodological issues will be discussed.

A. Feasibility

In order to evaluate the feasibility of the approach, several usage-related aspects (study participation, usage frequency, recording duration) and, concerning the SotD, rather functionality-related aspects (accuracy of word or face recognition, recognized emotions, congruence of verbal and facial emotion recognition) were assessed.

1. Usage-related Aspects

There were individual differences regarding the study participation and the usage frequency of the examined features. Three patients participated longer than proposed. Especially one participant was motivated to exceed the estimated study duration by seven weeks. Two of the participants were not able to participate for the originally intended duration of twelve weeks, because they were recruited only eight weeks before the project-intern assessment period ended. Consequently, no patient terminated the participation prematurely, which might indicate an overall feasible, ambulatory system. In reference to the specific features, the SA was the most feasible. Four out of five participants used it almost on a daily basis or even more frequently with multiple assessments on single days. Due to their well-established role in other monitoring approaches for bipolar disorder with comparable response rates [40], [41], this result was expected. The TPA were only used by related parties of two participants. Solely in one case there were multiple entries. Therefore, within the current study setting, this feature was not feasible. This result could be explained within the context of the often-strained relationships of patients with bipolar disorders [21] or of the relationship status of the participants. Yet, another important issue might have been the web-based study approach that did not allow for push messages or reminders and might have been more effortful in its handling. In the future, this feature should be realized with associated smartphone applications. Thus modified TPA might be valuable and feasible with patients in reliable relationships and should be examined in larger studies, possibly conceptualized as family studies [42].

The SotD assessments were more effortful than the SA due to the installation of external study microphones and holders and because of the necessity to narrate a freely chosen story. Two of the participants still used the SotD on a regular basis (i.e., almost daily or almost every second day). The other three participants used the SotD much less frequently (i.e., on about ten to 16 percent of their participation days). Although these results show that the SotD (in its current version) might be feasible with some patients, modifications regarding its handling or the usage experience may be necessary to increase its general feasibility. The presented results further indicate that not only the SA but even more the SotD approach may be more feasible with younger patients. This confirms findings of earlier studies examining how the age of patients with bipolar disorder affects their new media usage [36], [37].

2. Functionality of the Emotion Recognition Approach

Following the insights of an earlier pre-study with healthy participants, the SotD study approach was extended by the external microphones and holders. This hardware change did improve the previously worse recording quality of the auditive information and, thereby, did increase the accuracy of the automatic transcriptions or the count of correctly recognized words to an acceptable level [43]. In concern to the video analysis, there were six assessments without working face recognition. The corresponding recordings were examined more closely and could be explained with unfavorable recording angles and a relatively low lighting. As a consequence, instructions should be even more specific. Furthermore, future systems should inform their users on missing face recognition or they should be trained to be less vulnerable for disturbances. By analyzing the expressions within both sources (auditive and visual), the recognition software found far more positive than negative emotions. Although several factors may have contributed to these results (see following sections), the recognized emotions might still reflect the expressed emotions of the participants: In order to explore further research questions, additional clinical data of the five study-patients was assessed. Most of them showed relatively stable mood throughout their study participation, including moderate (hypo) manic or mild depressive symptoms but no severe depressive episodes. Therefore, strong or persisting changes in their emotional experiences, that could have elevated the amount of expressed sadness during the SotD assessments, may not have been induced by psychopathological symptoms. Beyond that, most of the correlations regarding the emotional LIWC [38] or FACS [39] categories seem to support the consistency of the auditive and visual measures. At first glance, however, the strong negative correlations between the FACS parameters for happiness and the LIWC category of positive emotions with one participant seem conflicting.

Clinically speaking, these results could still be explained: The participant showed mild depressive symptoms and a limited emotional reagibility throughout the whole study but yet parathymic smiles during the bi-weekly clinical assessments. These deficits coincide with findings of earlier studies indicating difficulties of patients with bipolar disorders in the (facial) expression of negative emotions [13], [14]. Therefore, the emotion recognition results presented here might represent a further indication of disease-specific deficits in the emotional expressions of bipolar patients. The verbally expressed information could have consciously been adapted to the assessment situation, while the missing correspondence within the facial expressions might have represented a more basic process of emotional avoidance. This interpretation might, in part, also explain the higher percentage of positive emotions within the FACS data of all patients (as compared to the LIWC results). Nevertheless, technical or conceptual issues may have further contributed to these results: Whereas the LIWC [38] and FACS [39] approaches delivered an empirically established framework for emotion recognition, the FACS-based video analysis might have been less sensitive to the recognition of negative emotions due to insufficient training data during the aforementioned pre-study.

Moreover, the results might suggest that short recordings of approximately two minutes or less could contain sparse emotional expressions altogether. In that case, ambient sound samples or random voice features, which have shown some potential regarding the prediction of mood state changes in bipolar disorder [27], [44]–[47], could increase the obtained information. However, as far as we know, none of the existing mobile-based approaches for bipolar disorder have analyzed the emotional content of verbal or facial expressions. It would already be difficult to realize the emotion recognition approach with random sound samples. The assessment of random video data with the emotion-sensitive approach would almost be impossible without harming the (perceived) privacy of patients. With this in mind, the SotD approach goes beyond the existing mHealth systems for bipolar disorder [34]. It allows for the analysis of even more sensitive and personal ambulatory data without harming privacy issues.

B. Acceptability

In order to evaluate the acceptability of the approach, semi-structured interviews were conducted with each participant. In accordance with earlier studies that indicate the positive attitudes of bipolar patients towards technical assistance [35], [41], [48], the participants of the pilot study perceived all three examined system features positively. As for the TPA, although only two patients included other persons into the data acquisition, the participants emphasized the importance of this feature. This coincides with comparable opinions of patients with bipolar disorder during an earlier, project-related interview study [35]. Considering the discussed and not yet perfectly solved issues of the SotD module, the positive response of the participants on this feature is quite impressive. Although only two out of five study participants used the SotD regularly (about three to six times per week), all of them were open to a more regular usage. The mentioned conditions, like an easier usage experience (e.g., without microphone or holder) or automatic reactions towards the told stories, matched the original concept of the assistance system [34]. In accordance with the process of development, these aspects had not been implemented within the study setting. They should, however, be technically convertible in the future. Thus, from the acceptability point of view, the ambulatory long-term assessment of emotional expressions in bipolar disorder should be possible.

C. General Discussion

Overall, the results of the pilot study regarding the emotion recognition approach are promising (while the TPA were not feasible). The ambulatory study setting with participation times of several weeks increased the knowledge gain regarding the acceptability and feasibility of the approach by accounting for a long-term, realistic and natural environment [49], [50]. The results indicate that emotion-sensitive systems may be feasible and well-accepted, especially with younger patients. These findings coincide with the good feasibility or acceptability of mHealth systems using self-monitoring, sensor or wearable data with bipolar patients [25], [29], [30], [40] and with the positive attitudes of this patient group towards innovative, technical strategies for disease management [35].

The explorative approach of the pilot study with five patients, who were mostly considered as single cases, allowed for detailed insights into an ambulatory application scenario and into personal usage experiences. Although the SotD module "solely" requires its regular usage to assess active and passive emotion-related data, it comes along with more effort than common SA. Consequently, younger patients, who show a higher technical affinity [36], [37], used the study version of the SotD more frequently than older participants. And all participants were very specific about the conditions of a regular (e.g., several times per week) long-term usage. In accordance with this, future developments should be more practical (e.g., without microphones or holders), less irritating (no selfie mode) and should realize a perceived system-interaction or feedback during the SotD assessments. Furthermore, in case of the informed consent of patients and after thorough consideration of all ethical and legal implications, future systems could include the attending physicians or therapists. Thus, individual and sudden changes in the emotional expressions of patients could be thoroughly reflected during subsequent sessions.

Of course, the current approach and the small sample size with individual differences within the participation times of the patients limit the generalizability of the findings. Beyond that, detailed cost-benefit analyses would be necessary before implementing emotion-sensitive modules into disease management approaches for bipolar disorder. Therefore, more research is needed and, aside from its acceptability and feasibility, the mobile-based emotion recognition approach would have to provide valid and reliable results regarding all relevant emotions of mood episodes in bipolar disorder [3], [51]. As a consequence, it could facilitate the monitoring and understanding of emotional aspects in this disease and enable following research to examine its potential clinical value or contribution towards mood state recognition. The strong relation between emotional deficits of bipolar patients and their global and social functioning [4], [6]–[9], [13], [52] might make the gained information helpful to therapeutic approaches [13] and relapse prevention. Last but not least, a well-functioning, mobile-based emotion recognition approach could help our understanding of emotional experiences or expressions and their ramifications in other disorders as well.

V. Conclusion

As far as we know, the examined assistance system incorporates the first, mobile-based emotion recognition approach for bipolar disorder. Whereas the openness of patient groups towards technical or mobile-based assistance in their disease management has been investigated on several occasions, the pilot study shows that even the ambulatory assessment of audio and video data may be well-accepted and feasible. Beyond that, the approach allows for the long-term analysis of verbally and facially expressed emotions without harming the perceived privacy of patients or data privacy. Thus, the emotion-sensitive mHealth approach could affect other research areas or fields of application as well. However, to that end, some methodological and technical issues have to be addressed by future developments, and further empirical studies with larger samples of patients are necessary to increase the generalizability of the results.

Author Statement

The EmAsIn project was approved by the ethical committee of the Faculty of Behavioural and Cultural Studies, Ruprecht-Karls-University Heidelberg.

Acknowledgment

We would like to thank the participants of our study and the technical partners of our research project EmAsIn for their support. This work was supported by the German Federal Ministry of Education and Research [grant number 16SV7357].

Conflicts of Interest

The authors herewith declare no potential conflict of interest in respect to research, authorship and/or publication of this article.

References

[1] I. Grande, M. Berk, B. Birmaher, and E. Vieta, "Bipolar disorder," *Lancet*, vol. 387, no. 10027. pp. 1561–1572, 2016.

[2] J. V. Pinto *et al.*, "Remission and recurrence in bipolar disorder: The data from health outcomes and patient evaluations in bipolar disorder (HOPE-BD) study," *Journal of Affective Disorders*, vol. 268, pp. 150–157, 2020.

[3] L. A. Carolan and M. J. Power, "What Basic Emotions Are Experienced in Bipolar Disorder?," *Clinical Psychology & Psychotherapy*, vol. 18, no. 5, pp. 366–378, 2011.

[4] M. Paris, Y. Mahajan, J. Kim, and T. Meade, "Emotional speech processing deficits in bipolar disorder: The role of mismatch negativity and P3a," *Journal of Affective Disorders*, vol. 234, pp. 261–269, 2018.

[5] A. C. Bilderbeck *et al.*, "Associations between mood instability and emotional processing in a large cohort of bipolar patients," *Psychological Medicine*, vol. 46, no. 15, pp. 3151–3160, 2016.

[6] S. L. Johnson, C. S. Carver, and J. A. Tharp, "Suicidality in Bipolar Disorder: The Role of Emotion-Triggered Impulsivity," *Suicide and Life-Threatening Behavior*, vol. 47, no. 2, pp. 177–192, 2017.

[7] S. L. Johnson *et al.*, "Emotion in Bipolar I Disorder: Implications for Functional and Symptom Outcomes," *Journal of Abnormal Psychology and Symptom Outcomes*, vol. 125, no. 1, pp. 40–52, 2015.

[8] A. Aparicio, J. L. Santos, E. Jiménez-López, A. Bagney, R. Rodríguez-Jiménez, and E. M. Sánchez-Morla, "Emotion processing and psychosocial functioning in euthymic bipolar disorder," *Acta Psychiatrica Scandinavica*, vol. 135, no. 4, pp. 339–350, 2017.

[9] L. D. Branco, C. Cotrena, A. Ponsoni, R. Salvador-Silva, S. J. L. Vasconcellos, and R. P. Fonseca, "Identification and Perceived Intensity of Facial Expressions of Emotion in Bipolar Disorder and Major Depression," *Archives of Clinical Neuropsychology*, vol. 33, no. 4, pp. 491–501, 2018.

[10] J. Gray *et al.*, "Bipolar patients show mood-congruent biases in sensitivity to facial expressions of emotion when exhibiting depressed symptoms, but not when exhibiting manic symptoms," *Cognitive Neuropsychiatry*, vol. 11, no. 6, pp. 505–520, 2006.

[11] H. R. Venn *et al.*, "Perception of facial expressions of emotion in bipolar disorder," *Bipolar Disorders*, vol. 6, no. 4, pp. 286–293, 2004.

[12] C. M. Hoertnagl *et al.*, "Combined processing of facial and vocal emotion in remitted patients with bipolar i disorder," *Journal of the International Neuropsychological Society*, vol. 25, no. 3, pp. 275–284, 2019.

[13] G. Bersani *et al.*, "Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: A partially shared cognitive and social deficit of the two disorders," *Neuropsychiatric Disease and Treatment*, vol. 9, pp. 1137–1144, 2013.

[14] I. Broch-Due, H. L. Kjærstad, L. V. Kessing, and K. Miskowiak, "Subtle behavioural responses during negative emotion reactivity and down-regulation in bipolar disorder: A facial expression and eye-tracking study," *Psychiatry Research*, vol. 266, pp. 152–159, 2018.

[15] R. P. Bentall, P. Kinderman, and K. Manson, "Self-discrepancies in bipolar disorder: Comparison of manic, depressed, remitted and normal participants," *British Journal of Clinical Psychology*, vol. 44, no. 4, pp. 457–473, 2005.

[16] M. L. Inder, M. T. Crowe, P. R. Joyce, S. Moor, J. D. Carter, and S. E. Luty, "'I really don't know whether it is still there': Ambivalent acceptance of a diagnosis of bipolar disorder," *Psychiatric Quarterly*, vol. 81, no. 2, pp. 157–165, 2010.

[17] L. M. Weinstock, T. Chou, C. Celis-deHoyos, I. W. Miller, and J. Gruber, "Reward and Punishment Sensitivity and Emotion Regulation Processes

Differentiate Bipolar and Unipolar Depression," *Cognitive Therapy and Research*, vol. 42, no. 6, pp. 794–802, 2018.

[18] J. Pech, M. Akhøj, J. Forman, L. V. Kessing, and U. Knorr, "The impact of a new affective episode on psychosocial functioning, quality of life and perceived stress in newly diagnosed patients with bipolar disorder: A prospective one-year case-control study," *Journal of Affective Disorders*, vol. 277, pp. 486–494, 2020.

[19] F. Bennett *et al.*, "Predictors of psychosocial outcome of bipolar disorder: data from the Stanley Foundation Bipolar Network," *International Journal of Bipolar Disorders*, vol. 7, no. 1, 2019.

[20] A. López-Villarreal *et al.*, "Progression of the functional deficit in a group of patients with bipolar disorder: a cluster analysis based on longitudinal data," *European Archives of Psychiatry and Clinical Neuroscience,* vol. 270, no. 8, pp. 947–957, 2020.

[21] M. J. Gitlin and D. J. Miklowitz, "The difficult lives of individuals with bipolar disorder: A review of functional outcomes and their implications for treatment," *Journal of Affective Disorders*, vol. 209, no. July 2016, pp. 147–154, 2017.

[22] C. J. L. Murray *et al.*, "Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010," *Lancet*, vol. 380, no. 9859, pp. 2197–2223, 2012.

[23] S. Miziou *et al.*, "Psychosocial treatment and interventions for bipolar disorder: a systematic review.," *Annals of General Psychiatry*, vol. 14, p. 19, 2015.

[24] E. Morton, E. E. Michalak, R. Hole, S. Buzwell, and G. Murray, "'Taking back the reins' – A qualitative study of the meaning and experience of self-management in bipolar disorder," *Journal of Affective Disorders*, vol. 228, pp. 160–165, 2018.

[25] D. Hidalgo-Mazzei, A. Mateu, M. Reinares, A. Matic, E. Vieta, and F. Colom, "Internet-based psychological interventions for bipolar disorder: Review of the present and insights into the future," *Journal of Affective Disorders*, vol. 188, pp. 1–13, 2015.

[26] E. Gliddon, S. J. Barnes, G. Murray, and E. E. Michalak, "Online and mobile technologies for self-management in bipolar disorder: A systematic review.," *Psychiatric Rehabilitation Journal*, vol. 40, no. 3, pp. 309–319, 2017.

[27] M. Matthews *et al.*, "Development and Evaluation of a Smartphone-Based Measure of Social Rhythms for Bipolar Disorder," *Assessment*, vol. 23, pp. 472–483, 2016.

[28] C. A. Depp *et al.*, "Augmenting psychoeducation with a mobile intervention for bipolar disorder: A randomized controlled trial," *Journal of Affective Disorders*, vol. 174, pp. 23–30, 2015.

[29] D. Hidalgo-Mazzei *et al.*, "OpenSIMPLe: A real-world implementation feasibility study of a smartphone-based psychoeducation programme for bipolar disorder," *Journal of Affective Disorders*, vol. 241, pp. 436–445, 2018.

[30] M. Faurholt-Jepsen *et al.*, "Smartphone-based self-monitoring in bipolar disorder: evaluation of usability and feasibility of two systems," *International Journal of Bipolar Disorders*, vol. 7, no. 1, pp. 1–11, 2019.

[31] E. Mühlbauer *et al.*, "Effectiveness of smartphone-based ambulatory assessment (SBAA-BD) including a predicting system for upcoming episodes in the long-term treatment of patients with bipolar disorders: Study protocol for a randomized controlled single-blind trial 11 Medical a," *BMC Psychiatry*, vol. 18, no. 1, 2018.

[32] J. Zulueta *et al.*, "Predicting mood disturbance severity with mobile phone keystroke metadata: A biaffect digital phenotyping study," *Journal of Medical Internet Research*, vol. 20, no. 7, pp. 1–10, 2018.

[33] A. Cochran, L. Belman-Wells, and M. McInnis, "Engagement strategies for self-monitoring symptoms of bipolar disorder with mobile and wearable technology: Protocol for a randomized controlled trial," *Journal of Medical Internet Research*, vol. 20, no. 5, 2018.

[34] H. Daus, T. Bloecher, R. Egeler, R. De Klerk, W. Stork, and M. Backenstrass, "Development of an emotion-sensitive mobile Health approach for mood state recognition in bipolar disorder," *JMIR Mental Health*, vol. 7, no. 7, pp. 1–10, 2020.

[35] H. Daus, N. Kislicyn, S. Heuer, and M. Backenstrass, "Disease management apps and technical assistance systems for bipolar disorder_ Investigating the patients´ point of view," *Journal of Affective Disorders*, vol. 229, pp. 351–357, 2018.

[36] R. Bauer *et al.*, "Internet use by older adults with bipolar disorder," *International Journal of Bipolar Disorders*, vol. 6, no. 1, p. 20, 2018.

[37] R. Bauer *et al.*, "Internet use by patients with bipolar disorder: Results from an international multisite survey," *Psychiatry Research*, vol. 242, 2016.

[38] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic Inquiry and Word Count," *Word Journal Of The International Linguistic Association.* pp. 1–21, 2001.

[39] P. Ekman and W. V. Friesen, "Manual for the facial action coding system," *Consulting Psychologists Press*, 1978.

[40] S. Schwartz, S. Schultz, A. Reider, and E. F. H. Saunders, "Daily mood monitoring of symptoms using smartphones in bipolar disorder: A pilot study assessing the feasibility of ecological momentary assessment," *Journal of Affective Disorders*, vol. 191, pp. 88–93, 2016.

[41] D. Hidalgo-Mazzei *et al.*, "Psychoeducation in bipolar disorder with a SIMPLe smartphone application: Feasibility, acceptability and satisfaction," *Journal of Affective Disorders*, vol. 200, pp. 58–66, 2016.

[42] T. J. Rothausen, "'Family' in organizational research: A review and comparison of definitions and measures," *Journal of Organizational Behavior*, vol. 20, no. 6, pp. 817–836, 1999.

[43] M. M. Al-Aynati and K. A. Chorneyko, "Comparison of voice-automated transcription and human transcription in generating pathology reports," *Archives of Pathology & Laboratory Medicine*, vol. 127, no. 6, pp. 721–725, 2003.

[44] S. Abdullah, M. Matthews, E. Frank, G. Doherty, G. Gay, and T. Choudhury, "Automatic detection of social rhythms in bipolar disorder," *Journal of the American Medical Informatics Association*, vol. 23, no. 3, pp. 538–543, 2016.

[45] F. Or, J. Torous, and J.-P. Onnela, "High potential but limited evidence: Using voice data from smartphones to monitor and diagnose mood disorders.," *Psychiatric Rehabilitation Journal*, vol. 40, no. 3, pp. 320–324, 2017.

[46] M. Faurholt-Jepsen *et al.*, "Voice analysis as an objective state marker in bipolar disorder," *Translational Psychiatry*, vol. 6, p. e856, 2016.

[47] Z. Pan, C. Gui, J. Zhang, J. Zhu, and D. Cui, "Detecting Manic State of Bipolar Disorder Based on Support Vector Machine and Gaussian Mixture Model Using Spontaneous Speech," *Psychiatry Investigation*, vol. 15, no. 7, pp. 695–700, 2018.

[48] K. E. A. Saunders, A. C. Bilderbeck, P. Panchal, L. Z. Atkinson, J. R. Geddes, and G. M. Goodwin, "Experiences of remote mood and activity monitoring in bipolar disorder: A qualitative study," *European Psychiatry*, vol. 41, pp. 115–121, 2017.

[49] T. J. Trull and U. Ebner-Priemer, "Ambulatory assessment," *Annual Review of Clinical Psychology*, vol. 9, pp. 151–176, 2013.

[50] U. W. Ebner-Priemer and T. J. Trull, "Ecological Momentary Assessment of Mood Disorders and Mood Dysregulation," *Psychological Assessment*, vol. 21, no. 4, pp. 463–475, 2009.

[51] American Psychiatric Association, "Diagnostic and statistical manual of mental disorders (5th ed.)," *American Psychiatric Publishing*, 2013.

[52] I. M. M. Lima, A. D. Peckham, and S. L. Johnson, "Cognitive deficits in bipolar disorders: Implications for emotion," *Clinical Psychology Review*, vol. 59. pp. 126–136, 2018.

Henning Daus

Henning Daus is psychologist and psychotherapist (CBT). He graduated from Eberhard Karls University Tübingen and worked as research associate at the Insitute of Clinical Psycholgy at the Stuttgart Hospital. His research interests include affective disorders, psychotherapy research or mHealth developments in bipolar disorder.

Matthias Backenstrass

Matthias Backenstrass is currently head of the Institute of Clinical Psychology at the Stuttgart Hospital. He is also Professor at the Institute of Psychology, University of Heidelberg. His research interests include affective disorders, especially bipolar disorder and persistent depressive disorder, emotion regulation, and psychotherapy research. He is psychologist and psychotherapist (CBT).

A Case-Based Reasoning Model Powered by Deep Learning for Radiology Report Recommendation

Elvira Amador-Domínguez[1], Emilio Serrano[1], Daniel Manrique[2], Javier Bajo[1] *

[1] Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid (Spain)
[2] Laboratorio de Inteligencia Artificial, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid (Spain)

Received 8 February 2021 | Accepted 29 May 2021 | Published 11 August 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

Case-Based Reasoning models are one of the most used reasoning paradigms in expert-knowledge-driven areas. One of the most prominent fields of use of these systems is the medical sector, where explainable models are required. However, these models are considerably reliant on user input and the introduction of relevant curated data. Deep learning approaches offer an analogous solution, where user input is not required. This paper proposes a hybrid Case-Based Reasoning, Deep Learning framework for medical-related applications, focusing on the generation of medical reports. The proposal combines the explainability and user-focused approach of case-based reasoning models with the deep learning techniques performance. Moreover, the framework is fully modular to fit a wide variety of tasks and data, such as real-time sensor captured data, images, or text, to name a few. An implementation of the proposed framework focusing on radiology report generation assistance is provided. This implementation is used to evaluate the proposal, showing that it can provide meaningful and accurate corrections, even when the amount of information available is minimal. Additional tests on the optimization degree of the case base are also performed, evidencing how the proposed framework can optimize this base to achieve optimal performance.

Keywords

Case-Based Reasoning, Deep Learning, Natural Language Processing, Entity Recognition, Medical Radiology.

DOI: 10.9781/ijimai.2021.08.011

I. Introduction

Deep Learning is currently a fundamental approach in Artificial Intelligence applied to the medical domain. Their applications include image segmentation [1]–[3], 3D image reconstruction [4], [5], and disease diagnosis [6], [7]. While these approaches offer outstanding results, they suffer from a considerable flaw: lack of explainability. This issue is particularly concerning in the medical domain, where it is crucial to understand the inference procedure carried by a model to perform a task. Moreover, deep learning-based approaches require a considerable amount of labelled data to be truly accurate, which may not always be available.

Opposite to this approach, the Case-Based Reasoning (CBR) methodology [8], [9] provides computational models closely related to human reasoning. In CBR, the resolution of problems provides knowledge that permits to solve new, similar ones. A CBR model discovers the closest situation to the current one to solve and adapt its solution to fit the present scenario. One of CBR's essential advantages is that it is easy to follow and understand the inference process they conduct, which has prompted its use in, for example, the medical domain [10], [11].

This paper proposes a hybrid CBR-deep learning model to tackle the problem of radiology report writing assistance. The main efforts in the radiology domain reside within image-related tasks, such as diagnosis or X-ray image segmentation. In this image-dominated field, medical reports play a secondary role, mostly used to support the aforementioned tasks. Thus, high quality labelled textual data in this domain may not always be available, which hinders the use of deep learning techniques.

The proposed approach uses a CBR model to work with a few cases that can scale up, assisted by deep learning models to improve its performance. Therefore, it is a blended solution between a knowledge-based system [12], where the knowledge must be elicited, and a deep learning model, where no expert assistance is required. The proposed CBR model considers expert knowledge as an input to improve and validate the stored cases, but it does not rely exclusively on this knowledge to function.

This framework has been developed under a Horizon 2020 research project AI4EU [13], whose goal is to provide users with artificial intelligence resources that satisfy specific user necessities. Moreover, resources developed under this project should be explainable, verifiable, physical, collaborative, and integrative [14]. The proposed system meets all these specifications, as the usage of a CBR model ensures explainability, collaboration, and verification. The combination of different machine learning modules within the proposed model enables integration. Simultaneously, the introduction of sensor-retrieved and human-generated data ascertains physical interaction between

* Corresponding author.

E-mail addresses: eamador@fi.upm.es (E. Amador-Domínguez), emilioserra@fi.upm.es (E. Serrano), daniel.manrique@upm.es (D. Manrique), jbajo@fi.upm.es (J. Bajo).

the users and the framework. The implementation of the proposed framework for the radiology domain is available as a resource in the project platform with an open-source software license [15]. This implementation can be accessed by any user and modified accordingly to fit different purposes and work domains.

The remainder of the paper is organized as follows. Section II provides an insight into the related works. Section III presents the proposed hybrid CBR-deep learning model for radiology report recommendation, while the architecture and implementation of the model are explained in Section IV. Section V reports experimentation and obtained results. Finally, Section VI draws conclusions and future work.

II. Related Works

Case-Based Reasoning is widely used in the medical domain due to its adaptability and interpretability. CBR models have been successfully employed for diagnosis [10], [16], medical decision support [17], and patient monitorization [18], amongst other tasks.

CBR methodology [19] implements a continuous cycle, where the model improves over time by assimilating the knowledge acquired from the resolution of previous problems or cases. Subsequently, the model's performance relies on the number of stored cases and the relevance of those cases concerning the given situation. Mechanisms to efficiently store and manage the acquired knowledge are needed to reach an optimal case set. Several works have explored these issues, presenting new approaches for case retrieval and case-based maintenance.

Qin et al. [20] use heuristics to develop a new and efficient case retrieval algorithm. Daengdej et al. [21] study the substitution of the standard distance-based retrieval algorithm by a statistic-based method, focusing on the automobilist sector. Regarding case-based maintenance, Torrent-Fontbona et al. [11] present a model that combines case-based redundancy reduction with weight attribute learning to store and manage the cases efficiently. Nasiri et al. [22] explore the introduction of ontologies to manage and ensure the stored cases' semantic consistency.

Opposite to these naïve approaches, recent proposals aim to integrate deep learning techniques within the CBR cycle. As previously stated, while deep learning models are currently state of the art in most benchmarking tasks, their lack of explainability hinders their usage in the medical domain. Nonetheless, CBR methodologies can benefit from deep learning qualities by integrating them into different parts of the cycle. Such is the case of the work by Marie et al. [23], where they combine a CBR model with a Convolutional Neural Network to segment kidney radiographs. This proposal presents CBR as a solution to quality data insufficiency, serving as a preprocessing and augmentation mechanism for the network. Similarly, Corbat et al. [24] employ a combination of CBR with deep learning to efficiently segment medical images. Finally, Lamy et al. [25] study the possibility of exploiting CBR models' interpretability to explain the predictions of a deep neural network over a breast cancer dataset.

Other proposals focus on the introduction and management of ad-hoc captured data via sensors. The introduction of this data enables the development of several healthcare-related applications. Tang et al. [26] employ a CBR model to analyze sensor retrieved data from nursing homes to develop personalized healthcare plans for the patients. On the other hand, approaches such as Massie et al. [27] and Forbes [28] focus on patient monitorization and risk prevention, detecting potentially dangerous cases.

While Case-Based Reasoning models have been successfully employed for image-related tasks, including the radiology domain, their applications on textual data have been much less explored. Deep learning techniques are currently state of the art in most radiology-related tasks, such as medical text classification [29]–[31], diagnosis [32]–[34] and event detection [35], [36]. Some works explore the idea of assisting experts in the generation of medical reports. Toledo et al. [37] propose a prototype of web-based speech recognition for the construction of medical reports, while Donnelly et al. [38] evaluate the comparison between radiology free-text versus structured reports.

III. Deep Learning Supported Case-Based Reasoning for the Generation of Medical Reports

This work presents a CBR deep learning supported model to assist in the medical report generation task. The proposed framework does not automatically generate medical reports but serves as an assistant that provides formal corrections, references, and suggestions. Opposite to the methods studied in Section II, the case-based reasoning model is the core of the proposal. Besides, the user is actively involved in the system's learning procedure, determining which outputs are valid and not, directly impacting the learning process.

The proposed case-based reasoning framework comprises four stages in a cycle: retrieve, reuse, retain, and revise. Fig. 1 presents an overview of the model, showing its four cyclic phases, the interactions between them and the case set, and between the system and the user. The design of the framework is modular to make it easily customizable to fit different problems and domains. The figure depicts interchangeable elements as building blocks.

A. Retrieve

The cycle begins when the user introduces a new problem or case. A case can be either a simple draft of a medical report, or include additional information such as images, specific terms, or references. When the user introduces a new case into the system, the first step is to determine the closest ones from the existing case set. A naïve approach to this issue is to use a simple K-NN search, where the amount of desired cases to retrieve, *K*, is set, and the selection is purely based on distance criteria between samples. While this approach offers a straightforward, efficient solution, two main shortcomings hinder its usage for the proposed system. First, the input data is not measurable. Second, the usage domain is expert-oriented, so that it requires more specific, hand-crafted criteria to retrieve similar cases accurately.

While the similarity between medical reports can be measured according to specific metrics like the age of the patient or demographic data, there are no fixed, static criteria that enable direct comparison. Moreover, while some elements may remain stable between comparisons, some criteria may vary between users. The proposed framework includes an indicator-based retrieval algorithm to tackle this issue. Instead of comparing each case as a whole, the algorithm evaluates four different indicators per case. The four considered indicators I1, I2, I3, and I4 are:

- I1: *Image Comparison.* While images may be irrelevant in some medical areas, they are the cornerstone in others like neurology or dermatology. In such fields, pictures provide essential information that should not be diluted within the text but treated separately. Several methods can be considered for image comparison, ranging from histogram to feature vector comparison. While Convolutional Neural Networks are possibly the most robust way to represent images in a fixed dimensional space, some simpler alternatives can be considered for the task. Feature matching algorithms such as SURF [39], ORB [40], or KAZE [41] offer interpretable, easy to implement options. Nonetheless, these algorithms are quite sensitive to potential image failures such as light flashings and cannot capture finer-grained information. A possible solution

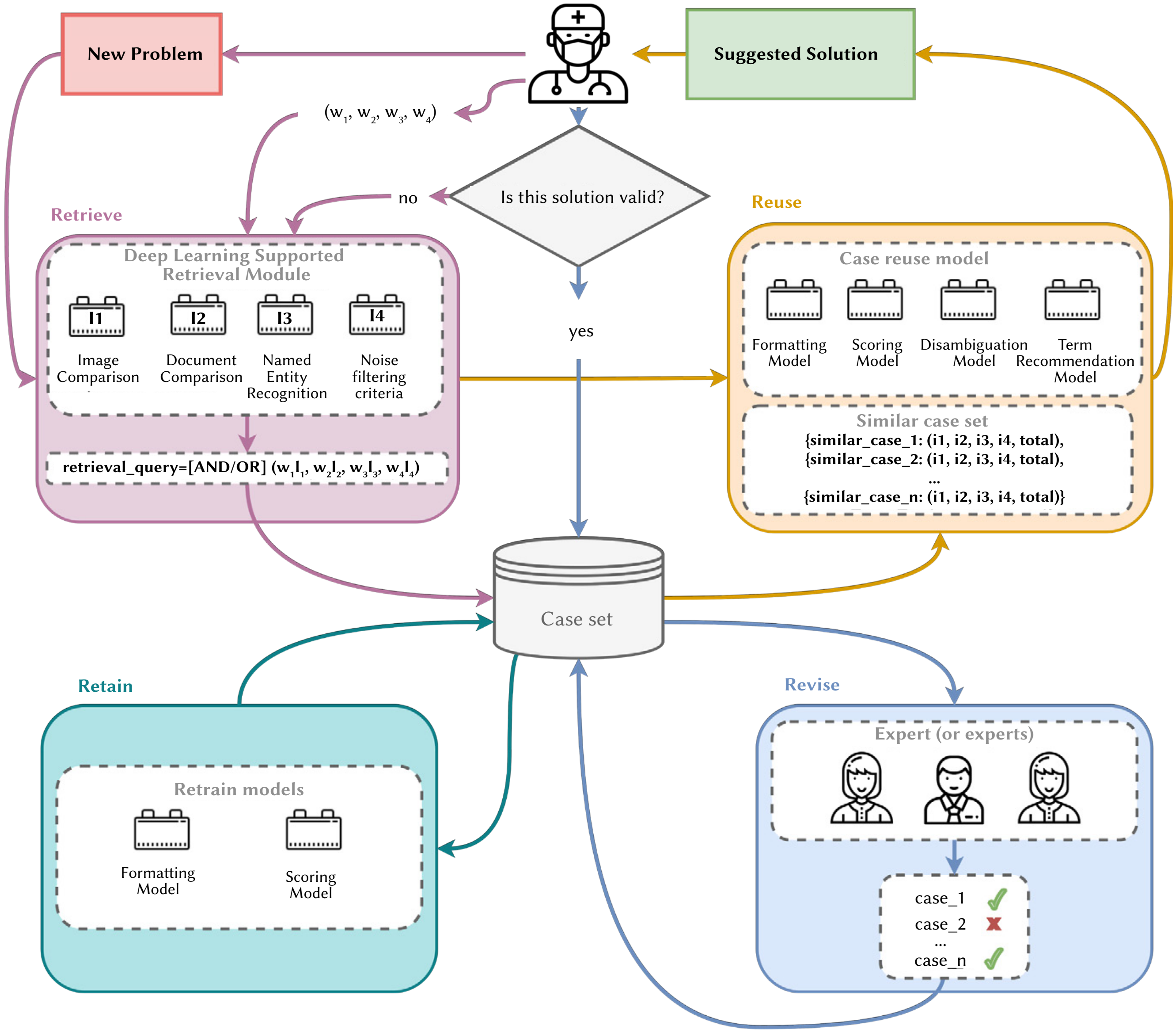


Fig. 1. An Overview of the Case-Based Reasoning System proposed.

to this issue is to combine differently generated feature vector representations into a unique vector. Once the image is embedded into a vector, a distance-based comparison can be established to ascertain the similarity between images.

- I2: *Document Comparison*. As in the case of images, several approaches can be considered to establish similarities between documents. While some non-feature-based methods can perform this task, their performance is entirely lacking compared to those where documents are embedded into a vector space and compared using different distance-based approaches. Models such as Word2Vec [42] or BERT [43] are the preferred choices for document representation, but more straightforward methods such as bag-of-words or TF-IDF can also be employed. However, these models cannot capture underlying semantic information, leading to less expressive representations at a faster cost. Regarding comparison, multiple methods can be considered depending on both the type of documents and the purpose. In this respect, cosine similarity, word's mover distance [44], or probabilistic based methods, which convert the embedding into a probabilistic distribution before comparison, are suitable choices.
- I3: *Named Entity Comparison*. Named Entity Recognition, or NER, is one of the main natural language processing tasks, particularly significant in the medical domain. In this task, the goal is to detect and label relevant terms within the text, such as people, places, or dates. While its usage is extended to a wide range of domains, there is a particular interest in developing NER models that focus specifically on detecting medical-related terms such as disease names, proteins, or drugs. Examples of clinical NER models are CliNER [45], BioBERT [46], or SciBERT [47]. Discovering relevant labelled terms within the documents is a way to detect and retrieve related documents. Therefore, only those cases whose reports contain user-specified terms will be considered for retrieval, reducing the search scope.
- I4: *Noise Filtering Criteria*. In addition to the previous indicators, additional filtering criteria may be specified to discard unfitting cases. They regard formatting specifications, like the absence of images or language employed, or type of content such as unidentified words like typos or abbreviations. Filtering criteria can be as restrictive as required.

For each of these four indicators, the user can establish a threshold value. Indicators can be combined either in a conjunctive or disjunctive way, depending on the user's goal. These criteria are then translated into a search query, which will then be used to retrieve the top N, or all, existing cases meeting the user-provided constraints.

```
query = { I1 >=0.85 , I2 >=0.7 , I3=['
    Pulmonary Disease ', 'Pneumonia '], I4
    =[lang= 'en ', identified_abbrv_rate >=
     0.9 ], N=5, operation= 'OR '}
```

Listing 1. Example of a retrieval query.

Listing 1 depicts an example of a retrieval query. According to this query, the system provides the user with the top 5 cases that either:

- Include images that are at least 85% similar to the given ones.
- Contain a clinical report that has a similarity of at least 70%.
- Contain the medical terms 'pulmonary disease' and 'pneumonia'.
- Are written in English, and at least 90% of the report's abbreviations have been disambiguated.

Cases that meet the retrieval criteria are ordered in decreasing order according to their cumulative similarity across the four indicators. Then, the top *N* cases demanded by the user are returned.

B. Reuse

Once the user has defined the retrieval criteria, the existing cases that fit the imposed constraints are selected and presented. A brief explanation of why each case has been chosen is also provided to maintain the system expressive and understandable. Providing information such as the similarity rates between the current case and the retrieved ones explains the system's decision process while giving further guidance to the user. From the instances retrieved in the previous stage, several operations are performed to obtain precise, expressive information that will aid the user in the report generation task. Fig. 1 shows four different modules in this stage to provide information to the user:

- *Formatting Module.* Readability is one of the most desirable features when it comes to any written report. It encompasses content matters, such as syntactic cohesion, and more straightforward format issues like proper paragraph and sectioning when required. In medical documents, while there may be differences from one domain to another, there is generally a fixed, basic structure to present the data. On a general view, a medical report comprises four main sections:
    - *Indication.* A brief introduction to the case, giving superficial information about the patient and the observed symptoms.
    - *Comparison.* References to previous existing reports of the given patient.
    - *Findings.* In-depth information about the potential causes of the symptoms, as well as additional observations about the patient.
    - *Impression.* Conclusions and diagnosis.

  A formatting module is included in this stage such that when a report in raw format is introduced, it can be adequately divided into paragraphs and sections.

- *Disambiguation Module.* Abbreviations are quite usual in the medical domain due to the existence of a high amount of complicated compound term names. However, while most of these abbreviations may be universal and easily understandable by any professional, some can still be obscure for a regular reader. A potential solution for this issue is to include a module that not only detects the abbreviations contained in the report but offers disambiguation suggestions for them. While this may extend the document, it also highly improves its readability, as it removes any potential misunderstandings induced by the abbreviations.
- *Term Recommendation Module.* As previously stated, Named Entity Recognition is particularly prominent in the medical domain. These models can accurately detect relevant terms and group them in a fixed set of given categories. These categories are usually related to each other in some manner, and, subsequently, so are the corresponding terms. For example, given a report about a patient with pulmonary disease, terms such as *(pneumonia, disease)* and *(chest x-ray, test)* may appear together frequently. This module offers these correlated terms to the user as suggestions. To obtain these suggested terms, named entity recognition is performed over the previously retrieved relevant cases, receiving a set of terms with their corresponding category. This set of terms are then flattened, cleaned, and presented to the user in their corresponding categories. Hence, if the user is writing a report containing the word *pneumonia*, but does not include *chest x-ray*, the system may recommend the inclusion of this term, as this correlation has previously appeared in those cases detected as related.
- *Scoring Module.* Finally, the system presents the user with a validity score, indicating whether the report, in its current form, is readable and understandable enough. This score can vary from a simple binary value (valid or invalid) to a star-scored base method, to a finer decimal system.

It is important to note that the recommendations and suggestions offered by the system are not final. The user must decide which of the given suggestions are to be applied to the current report.

Once the report's state satisfies the user, the system generates a new case and stores it into the case base. It also presents the original document as the problem and the final state as the solution. New cases are marked as *pending validation* and will not be added to the case base until the experts validate them.

C. Revise

Once the report satisfies the user, after applying any or none of the suggestions provided, the system generates a new case. However, it cannot be added directly to the case set as it may include errors that can hinder the system from improving. Moreover, if the system stored unreliable cases without any revision, they might be presented to the next users as solutions, misguiding them. Therefore, an intermediate step is required to ensure that those instances included in the case set are useful and needed.

A panel of experts must perform this task, manually checking *pending-on-validation* cases to provide them with a coherent score with the criteria implemented in the scoring module. Hence, if the system offers the user a binary score, the experts must also grade the cases following this criterium. Experts can also modify or correct minor mistakes within the cases before validating them to ensure their quality.

D. Retain

As noted in Section II, one of the biggest concerns regarding case-based reasoning systems is how to handle the ever-growing number of cases. Ideally, the case base should be composed of an optimal number of instances where the problem coverage is maximum, while the number of cases is minimum. However, while infeasible cases may not help the user, they improve the scoring models' accuracy. For this purpose, invalidated cases are also stored separately from the case base, where they can be recovered when necessary.

New cases are being regularly introduced into the case base and, subsequently, they must affect the system's behaviour. CBR models nurture themselves by adding further information, which keeps

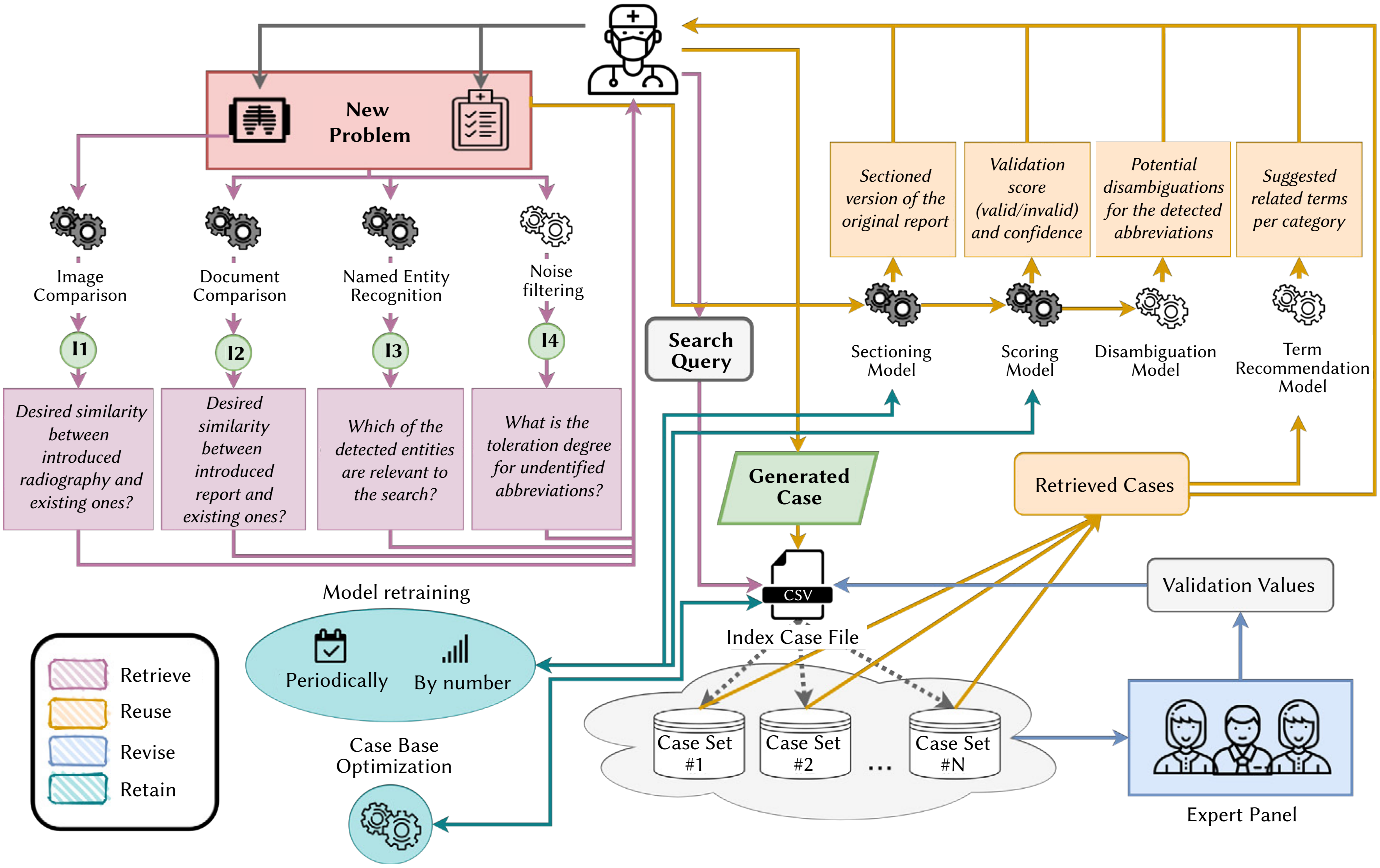

New Problem
Image Comparison
Document Comparison
Named Entity Recognition
Noise filtering
I1
I2
I3
I4
Desired similarity between introduced radiography and existing ones?
Desired similarity between introduced report and existing ones?
Which of the detected entities are relevant to the search?
What is the toleration degree for undentified abbreviations?
Search Query
Sectioned version of the original report
Validation score (valid/invalid) and confidence
Potential disambiguations for the detected abbreviations
Suggested related terms per category
Sectioning Model
Scoring Model
Disambiguation Model
Term Recommendation Model
Generated Case
Retrieved Cases
CSV
Index Case File
Validation Values
Model retraining
Periodically
By number
Case Base Optimization
Case Set #1
Case Set #2
...
Case Set #N
Expert Panel
Retrieve
Reuse
Revise
Retain

Fig. 2. Data flow of the proposed implementation. Coloured elements depict each stage of the CBR cycle. Solid gears represent deep learning powered modules, while clear ones represent non-machine learning modules.

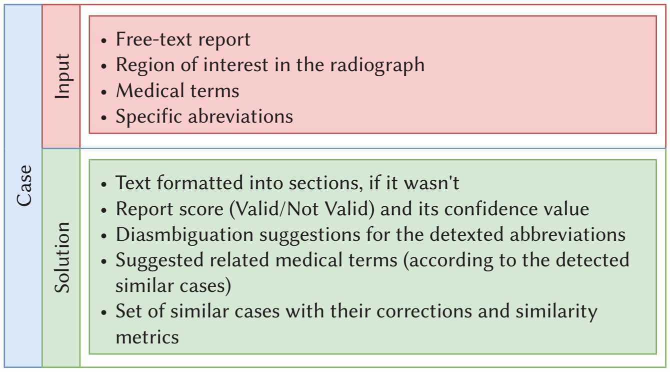
them updated and usable throughout time. Aside from case-based maintenance, module updates also are conducted in this stage. These updates can be either a replacement, such as switching from regular expressions to machine learning models, or just a retrain of an existing model. Updates can be either scheduled periodically or when a particular milestone of case numbers is reached. The scoring model can eventually substitute the panel of experts once it has gained enough maturity.

IV. System Architecture and Implementation

An implementation and case study is provided to illustrate the proposed framework. In this case study, the system focuses on the treatment and generation of radiology reports. This context presents a challenging scenario where both images and textual information are highly relevant to the problem. The implemented resource instantiates the proposal depicted in Fig. 1, selecting the appropriate paradigms for each of the eligible modules. Fig. 2 illustrates the data flow of the system.

The framework implements a four-layered software architecture. Before defining the CBR, some issues need to be addressed, such as data management and storage mechanisms. An indexed storage model is employed to deal with the ever-growing nature of the case set while still enabling fast retrieval. In the proposed storage system, cases are stored either in a distributed or centralized way and are referenced in an index file. The index file contains each case's location and its respective retrieval indicators to accelerate the retrieval process. Before starting the CBR cycle, preprocessing operations may be required to fit the system's constraints, such as separating images from text or formatting the report.

In the context of radiology, a case comprises a radiograph and a brief text summarizing the most relevant findings of the image, alongside additional information about the patient. While these two elements are enough to define a new problem, the user can also provide further information, as depicted in Fig. 3.



- Free-text report
- Region of interest in the radiograph
- Medical terms
- Specific abreviations

- Text formatted into sections, if it wasn't
- Report score (Valid/Not Valid) and its confidence value
- Diasmbiguation suggestions for the detexted abbreviations
- Suggested related medical terms (according to the detected similar cases)
- Set of similar cases with their corrections and similarity metrics

Fig. 3. Case composition of the provided implementation.

The retrieve stage begins once the user introduces a new problem into the system. Then, case indicators are then computed as follows:

- I1. *Image comparison:* In the current domain, images are black and white radiographs. Hence, there is not much variation between samples. A convolutional neural network generates the embeddings to capture the subtle differences between radiographs and enable an accurate comparison. A white-box feature detection algorithm is also employed to add a supplementary explainable level to the comparison. KAZE [41] generates fixed dimension descriptors from the key points detected in an image. These key points can be indicated in the picture, providing a visual explanation based on which the comparison is performed. KAZE representation is averaged with that obtained from the convolutional neural network. Then, it generates a unique embedding that combines both interpretable and abstract knowledge. The comparison is performed based on this final combined embedding.

1. I2. *Report comparison:* This task employs a pretrained NLP model specific to clinical data. This model provides single word embeddings for each of the tokens within the text, sentence-level embeddings, and document embeddings. The latest type is used to generate comparable report representations.
2. I3. *Named Entity Recognition:* This task uses CliNER [45]. This framework provides a series of models trained over a sizeable clinical corpus, capable of identifying the following entity types: diseases, treatments, and tests. As mentioned in Section III, multiple NER choices in the clinical domain range from fine-grained information, such as protein detection, to general type identification such as drugs versus diseases. CliNER offers an intermediate solution that fits the present scenario.
3. I4. *Noise filtering:* The same NLP model employed for report comparison is used to filter noise. In this context, noise refers to those elements on the text that can not be identified as tokens, and therefore they have no embedding nor meaning attached. The report is run through the NLP model to detect these conflicting terms, obtaining a set of identified tokens. Noise is then calculated as the proportion of identified tokens concerning the total amount of elements contained within the text.

These indicators are only computed once per case and are stored in the index file to accelerate the retrieval process. The user is then asked to specify which threshold values are considered for each of the proposed metrics, how to combine the indicators (conjunctively or disjunctively), and the number of related cases *k* which must be retrieved. Fig. 2 depicts a descriptive representation of the values inquired to the user, represented by purple-coloured boxes, where the threshold value for each indicator is posed as a human-readable question. For example, in the case of I2 (document processing), the framework would ask the user 'what is the minimum similarity acceptable between the current and the existing reports?'. These queries must be clearly presented and understandable to the user, as the success of the retrieval phase is directly related to the constructed query.

Once the search query is formulated, a comparison between the current problem and the existing cases is performed. Instead of retrieving each complete case individually from the case set, the comparison is performed based on the case indicators contained in the index file. Thus, when an existing case is detected as fitting, its full content is retrieved from the case set. A summary of each indicator's similarity metrics is attached and presented to the user alongside the case itself.

The retrieved cases are then used as a support for the term recommendation module This list containing the retrieved, top *k* similar cases is also provided to the user. Orange-coloured boxes in Fig. 2 present the different stages of the reuse phase. As shown, the named entities identified in the retrieved cases are processed by the term recommendation module, which groups the detected terms according to their type. Duplicate entries are also removed. The resulting term aggregations are then presented to the user, providing guidance on which entities could be related to the ones detected in the current case. Additionally, as depicted in Fig. 3, the following content and format suggestions are provided to the user as part of the solution:

- *Sectioned version of the report:* A bi-directional long-short term memory is employed for the formatting task. The problem itself is treated as a classification problem, where each sentence is labelled according to the section where it appeared. The goal of the model is to predict the best fitting section for each sentence. When formatting a new report, sentences are presented in the same order they are listed in the text to avoid permutations in the content.
- *Potential disambiguations for the detected abbreviations:* Similarly to the noise filtering operation, a set of unidentified tokens within the text is first obtained. The elements in this set are then looked up in the medical terminology SNOMED-CT, bringing the best applicable medical term for the input abbreviation.
- *Case validation score and confidence:* Binary scoring is employed in this implementation, categorizing the cases between valid and invalid. While a case is only validated or discarded in the revising stage, this score informs the user of whether the current state of the report would be considered appropriate or not. For this task, a random forest is used.
- *Suggested related terms per category:* Named entity recognition is applied to the content of the top *N* retrieved cases, obtaining a set of *(term, category)* tuples. Duplicates are removed from the set. These terms are then presented to the user grouped by category. CliNER [45] identifies named entities within the report, categorizing the detected terms into three types: disease, test, and treatment.

The system presents these suggestions to the user, who can freely decide which must be applied to the current problem. Once the appropriate modifications over the original report are performed, the generated solution is stored alongside the initial problem, comprising a new case. New cases are labelled as *pending on validation* and will not be shown to future users until experts have reviewed them.

During the phase of revise (depicted in Fig. 2 in blue-coloured boxes), an expert panel is in charge of regularly validating the pending cases, deciding which are valid and should be presented to the users and which are not. The validation status of each case is also referenced in the case index file to ease the filtering of which cases should be shown. Commonly, invalid cases are deleted from the case set, as they intuitively do not provide valuable information to the user. However, these cases are necessary to train and obtain robust scoring models that may even replace the expert at some point. Corrupted cases can be exploited for the benefit of the system, improving its performance.

Once there are enough classified cases, the retain stage begins, as depicted by the green-coloured bubbles in Fig. 2. In this stage, both the scoring and sectioning models employed in the reuse phase are retrained using the case set's information. Models can be retrained following either a periodical or a quantitative approach. Periodical retraining ensures that the model is kept updated and improves the final quality of the results. However, this approach presents a shortcoming: when there is a limited number of cases in the case set, the model's generalization capability will be logically limited. Additionally, case base optimization is performed in this stage. As previously stated, one of the biggest challenges in CBR models is to devise a management protocol for dealing with the ever-growing amount of cases. In the proposed framework, case base optimization is performed by maximizing case relation. First, a global linking process is launched amongst cases, computing the top 5 most similar cases per instance. Cases that are listed as related by at least one different case are kept in the case base. Unreferenced cases are removed from the case base, thus not shown to the users, but are still considered for model training.

V. Experimentation and Results

Experimentation based on the proposed implementation is set up to assess the performance and accuracy of the proposal. The majority of the studied approaches focus on evaluating the retrieval

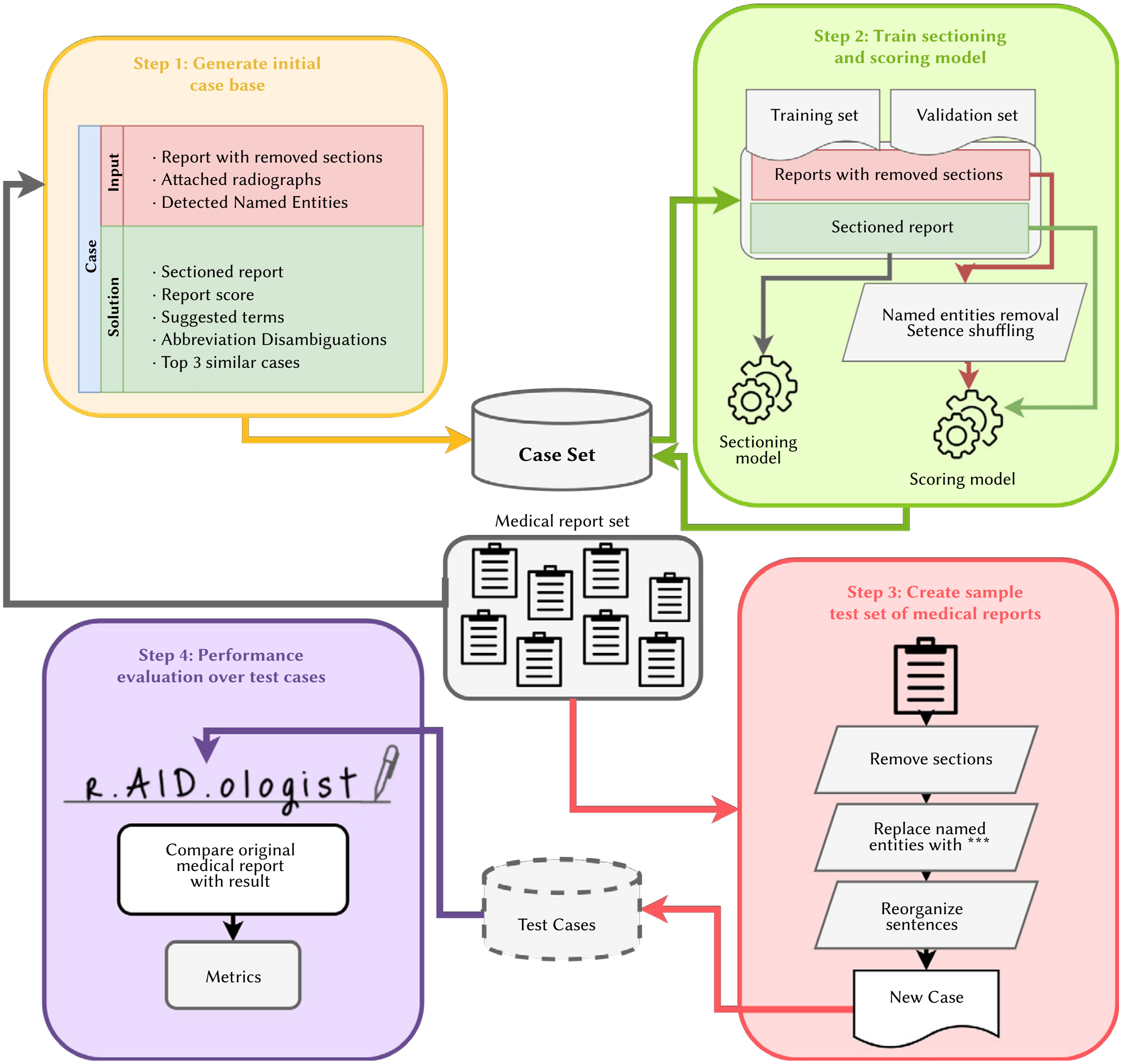
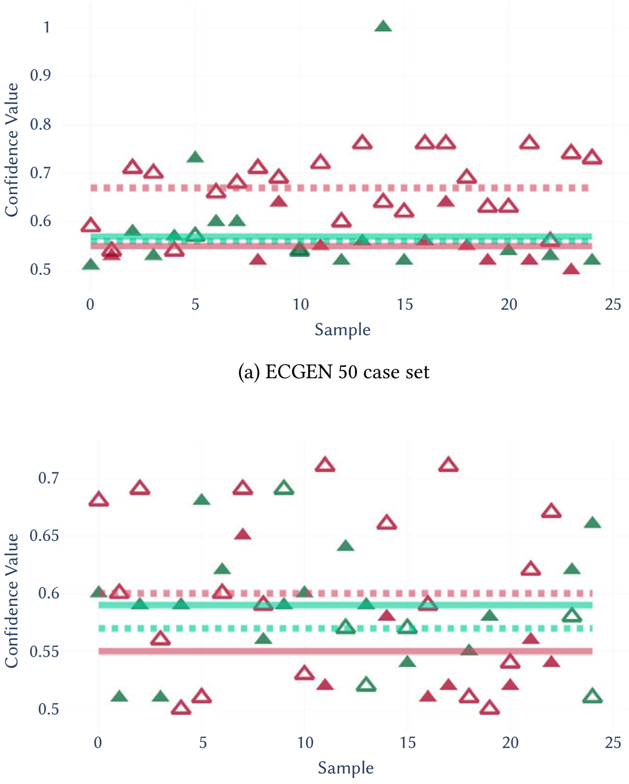

Step 1: Generate initial case base
Case
Input
· Report with removed sections
· Attached radiographs
· Detected Named Entities
Solution
· Sectioned report
· Report score
· Suggested terms
· Abbreviation Disambiguations
· Top 3 similar cases
Case Set
Step 2: Train sectioning and scoring model
Training set
Validation set
Reports with removed sections
Sectioned report
Named entities removal
Setence shuffling
Sectioning model
Scoring model
Medical report set
Step 3: Create sample test set of medical reports
Remove sections
Replace named entities with ***
Reorganize sentences
New Case
Test Cases
Step 4: Performance evaluation over test cases
R.AID.ologist
Compare original medical report with result
Metrics

Fig. 4. Overview of the experimentation process conducted to evaluate the system.

strategy, as it is a crucial element of CBR systems. Our proposal, however, relies on user-input queries to retrieve the most fitting cases. Hence, assessing the system performance based solely on the retrieval approach would not be representative enough, as the success of this stage is directly related to the user criteria.

Since the considered context is highly expert-oriented, it is not trivial to perform a quality assessment of the framework without expert information assistance. Therefore, an alternative evaluation approach capable of quantitatively measuring the performance of the model is required. The proposed evaluation procedure assesses the performance of the proposal for the report correction task. Fig. 4 depicts the conducted evaluation process, comprised of the following stages:

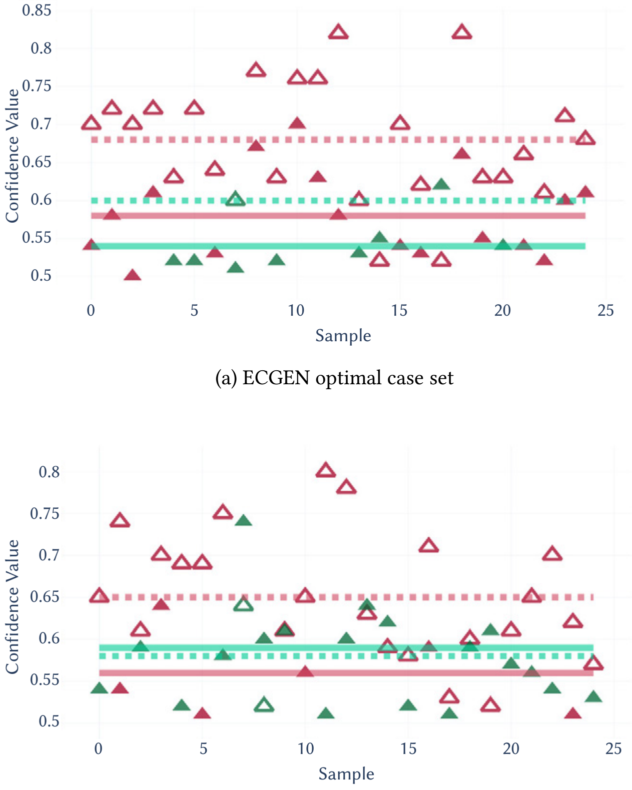
1. *Step 1: Generate the initial case base.* As previously stated, the case base is at the core of any case-based reasoning model. In this first step, a set of medical reports is converted into cases, composing the initial case base. Out of all the available medical reports, a sub-sample of 25 elements is randomly selected to be later used for testing. These randomly selected elements are not included in the case base. From the remaining cases, each medical report is stripped, when possible, from its sections, creating the input of the case. If a list of named entities and abbreviations are provided for the report, they are also included as the input. If the original report was already sectioned, its content is stored in the case solution as a sectioned report. The remaining solution values (score, suggested terms, and similar cases) are updated in the following step.
2. *Step 2: Train sectioning and scoring model.* At this stage, the cases contained in the case base only include the input (the original report stripped of its sections) and its corresponding solution (the original report without any modifications). These are the only attributes required to train both the sectioning and scoring model. The existing cases are randomly divided into two sets: training and validation. As stated in Section IV, sentence-based classification using a bi-directional long-short term memory is used to section each report. The sectioning model is trained using the case solution, where each report is split into sentences,



(a) ECGEN 50 case set

(b) MIMIC-CXR 50 case set

Fig. 5. Validation status on the 50-case-set for each dataset. Empty triangles denote the state of the case before the system applies the appropriate corrections. Solid triangles indicate their status afterward. Green and red colors depict valid and invalid cases, respectively. The y-axis represents the confidence value assigned by the system to the validation score.Average values are depicted as horizontal lines: discontinuous and continuous lines represent before and after values, respectively. The colors employed for the cases match the average lines.



(a) ECGEN optimal case set

(b) MIMIC-CXR optimal case set

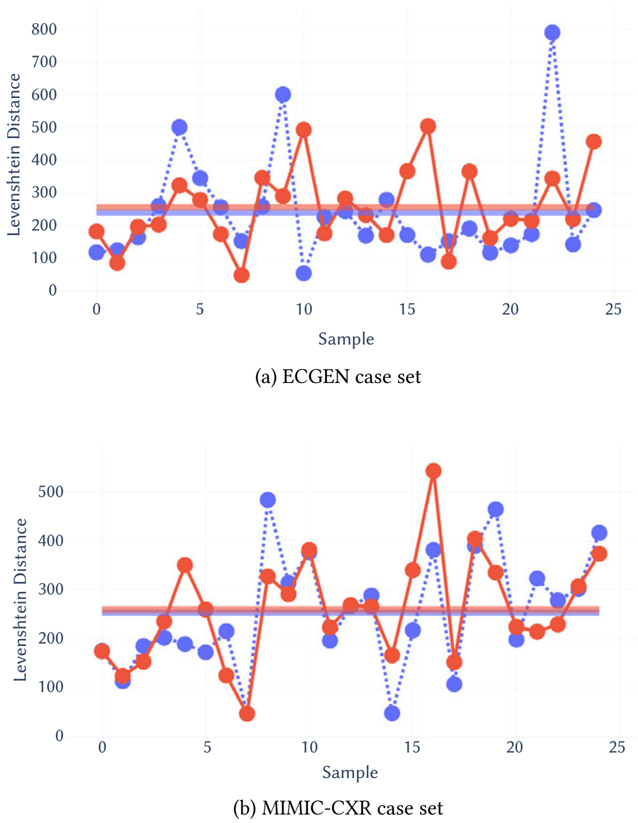
Fig. 6. Validation status on the optimal case base for each dataset. Empty triangles denote the state of the case before the system applies the appropriate corrections, while solid triangles denote their status afterward. Green and red colors depict valid and invalid cases, respectively. The y-axis represents the confidence value assigned by the system to the validation score. Average values are shown as horizontal lines: discontinuous and continuous lines illustrate before and after values, respectively. The same color code employed for the cases is used for the average lines.

and each sentence is labelled with the value of its corresponding section. For the scoring model, both the input and the solution of each case are required as this model feeds positive and negative samples. Therefore, case inputs comprise the negative sample set, while solutions comprise the positive sample set. A sequence of escape characters substitutes the named entities on each non-sectioned report, and the sentences are randomly reordered to further corrupt the negative samples. These sets are then used to train a random forest classifier, which acts as the scoring model. Once both sectioning and scoring models have been trained and validated, the case base is updated, adding each report score. Named entities, disambiguations, and similar cases are also updated.

3. *Step 3: Create a sample test set.* A set of input cases is created from the medical reports set aside for testing in Step 1. A comparison between the provided solution for a corrupted version of the input versus the original report is conducted to assess the proposal performance. Therefore, for each element in the test set, the following corruption operations are performed to create an input case: section removal, named entity replacement by a character sequence, and sentence reordering.
4. *Step 4: Performance evaluation over the test set.* The generated inputs are then passed onto the system, which attempts to provide a valid solution for the input permuted report. Alongside the corrected report, the framework presents a list of recommended terms and disambiguation abbreviations. The corrected version of the report is then compared with the original. The model should reorganize the sentences into sections in a cohesive order and suggest introducing the named entities previously stripped from the report. The following metrics are computed to assess the framework performance:
   (a) The validation score provided by the model before and after the corrections.
   (b) The Levenshtein distance between the original report and the suggested correction.
   (c) The proportion of entities detected on the original report pointed out by the model.

Two different radiology datasets are considered for evaluation: MIMIC-CXR [48] and Open-I's radiology set, denoted as ECGEN [49]. MIMIC-CXR contains complete medical reports in plain text format, without any additional information. On the contrary, Open-I provides both images and named entities alongside the medical report, and additional metadata. From each dataset, two initial case bases are generated, composed of 50 and 200 cases, respectively. Cases are generated from a random sampling of reports from each considered dataset. The developed implementation is used to conduct the experimentation.

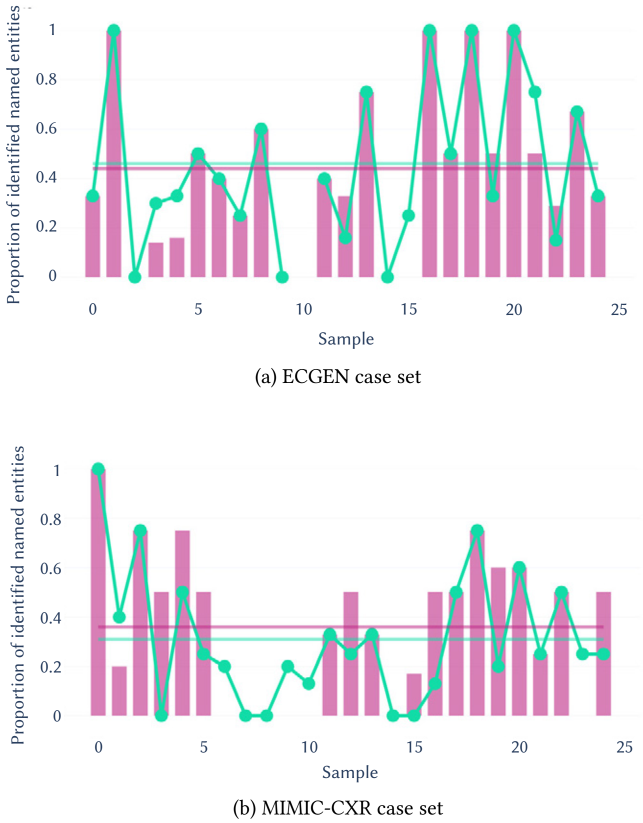
The initial 50-element case base serves as a baseline to assess the performance of the framework when the number of cases is limited. Applying the retain criteria in this scenario may not have any impact, as most or all cases may be related between them. In the initial



(a) ECGEN case set

(b) MIMIC-CXR case set

Fig. 7. Levenshtein distance per sample on each studied dataset. Purple and orange lines depict the results obtained on the 50-case and optimal set, respectively. The horizontal lines represent the average values, using the same color code employed for the results.



(a) ECGEN case set

(b) MIMIC-CXR case set

Fig. 8. Proportion of named entities correctly identified on each studied dataset. Purple bars depict the results obtained for the 50-case set, while green lines illustrate the results achieved in the optimal case base. The horizontal lines represent the average proportions, using the same color code employed for the results.

200-element case base, where the amount of existing elements is four times the size of the prior case base, retain criteria can be successfully applied, obtaining the optimal case base. The resulting optimal case bases are comprised of a total of 187 elements for MIMIC-CXR and 90 cases for ECGEN. Two different case bases are considered per dataset: a baseline 50-element case base, and an optimal case base.

Fig. 5 depicts the results obtained by the model when the case base comprises only 50 cases. Despite the simplicity of the case base, the framework still offers noteworthy results, accurately correcting most of the initially corrupted cases. This performance can be clearly observed in the results obtained in the ECGEN dataset (Fig. 5(a)), where most of the initial cases are noted as corrupted with a high confidence value and turn into valid after the corrections applied by the system. In the case of MIMIC-CXR, this improvement is not as noticeable as some cases remain considered invalid by the system after the corrections. However, as illustrated by Fig. 5(b), even in those cases still denoted as invalid after the modifications, the confidence value assigned by the system dramatically diminishes. This decrement evidence that, even though the report is still marked as invalid, the system corrections significantly reduce the corruption level of the report.

Using the optimal case set of each studied dataset impacts positively the performance of the model, as shown in Fig. 6. In this optimized context, the results are slightly more polarized than in the previous case, and most of the original corrupted cases are corrected and validated once processed by the system. Moreover, the confidence levels are higher than in the 50-case set, indicating that the framework can train more refined models, better distinguishing between valid and corrupt cases. Furthermore, considering the optimal case set for each particular dataset soothes the existing differences in performance. While in the 50-case set, the results obtained on the ECGEN dataset were slightly better since more reports were correctly modified and denoted as valid, in the MIMIC dataset the reports underwent a correction process that was insufficient to validate the case.

Levenshtein distance [50] between each original and corrected report pair is also computed to further assess the correction capabilities of the model. While different text similarity metrics could be considered for evaluation, such as cosine similarity or Jaccard index, these metrics do not consider text order. As previously stated, test cases are generated by stripping sections, permuting sentence order, and removing named entities. Therefore, even after the corruption process, both the original and corrupted report are almost equal in terms of content. Thus, an order-sensitive metric is required to ascertain the similarity degree between the original and corrected report. Fig. 7 illustrates the Levenshtein distance per pair of an original and corrected report on each studied dataset and case base. As shown in Fig. 7(a), Levenshtein distances in ECGEN, on both 50-element and optimal case bases, remains fairly similar throughout cases. A similar occurrence happens in MIMIC-CXR cases (Fig. 7(b)), where the distance between original and corrected reports remains akin. While finding the optimal case base benefited the framework results in the validation scenario, this improvement is not reflected regarding report sectioning and reordering. This flaw may be solved with the introduction of user input. While corrupted samples have been artificially generated from simple text editing operations in this experimentation, user-corrected reports may be more expressive and richer in content, leading the model to identify more complex correction patterns that would subsequently lead to better results.

The amount of named entities correctly suggested by the system is also provided, illustrated in Fig. 8. As stated in step 4 of the experimentation process, named entities in the original report are substituted by escape characters as part of the corruption process. Figures 8(a) and 8(b) depict the proportion of named entities stripped from the original report and correctly suggested by the framework on each dataset. The results show that, when the case base is optimized, the amount of detected entities either improves or holds. This is particularly noticeable in ECGEN's results (Fig. 8(a)). Only in three cases, the amount of detected entities slightly decreases with the optimized case base, but significantly improves in four other cases. In MIMIC-CXR (Fig. 8(b)), the results are not as consistent, which could be due to the difference in the case base size between both studied datasets. MIMIC-CXR has double the cases on its optimized version than ECGEN. Named entities are suggested based entirely on the top *k* most similar cases identified by the system. Hence, if the retrieved similar cases contain few named entities, this would directly impact the number of suggestions provided by the system. A way to overcome this issue is to increase the value of *k*.

VI. Conclusions and Future Work

This work presents a hybrid framework that combines a case-based reasoning system with several deep learning models to help health professionals generate medical reports. The proposed system is fully modular, making it effortlessly adaptable to several scenarios and heterogeneous data. A use case focusing on the development of radiology reports is provided to illustrate the proposal. An open-source implementation for this particular use case named r.AID.ologist is provided under the AI4EU platform. This implementation is used to assess the performance of the proposed framework. Two different radiology datasets are used: MIMIC-CXR and ECGEN. For each studied dataset, two different scenarios are considered: a baseline 50-element case base and an optimized case base. The results show that, even without external user validation, the system considerably benefits from optimizing the case base, as it increments its sensibility. Moreover, the results also evidence the robustness of the proposal even when the amount of available information is minimal, being capable of properly correct formatting errors while providing relevant suggestions, such as related terms or abbreviation disambiguations.

Acknowledgments

The authors thank the reviewers and editors for their valuable comments and suggestions, which have improved this paper. This project has received funding from the Horizon 2020 research and innovation programme of the European Union, under grant agreement No. 825619. This work has also been supported by the Autonomous Region of Madrid through the program CABAHLA-CM (GA No. P2018/TCS-4423) and by the "Universidad Politécnica de Madrid" under the program "Ayudas para Contratos Predoctorales para la Realización del Doctorado". The authors would like to thank Jérémy Clech and Guillaume Martial from NEHS DIGITAL for their support and numerous comments during the development of this work.

References

[1] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao,T. Zhang, S. Gao, J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2281–2292, 2019, doi: 10.1109/TMI.2019.2903562.

[2] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*, San Diego, CA, USA, 2019, pp. 225–2255.

[3] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling*, New York, New York, USA, 2020, pp. 451–462, Springer International Publishing.

[4] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, M. Niessner, "Scan2cad: Learning cad model alignment in rgb-d scans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, June 2019.

[5] B. Yang, S. Wang, A. Markham, N. Trigoni, "Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction," *International Journal of Computer Vision*, vol. 128, pp. 53–73, Jan 2020, doi: 10.1007/s11263-019-01217-w.

[6] J. Liu, Z. Zhang, N. Razavian, "Deep ehr: Chronic disease prediction using medical notes," in *Proceedings of the 3rd Machine Learning for Healthcare Conference*, vol. 85 of *Proceedings of Machine Learning Research*, Palo Alto, California, 17–18 Aug 2018, pp. 440–464, PMLR.

[7] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, D. Rueckert, "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease," *Medical Image Analysis*, vol. 48, p. 117–130, Aug 2018, doi: 10.1016/j.media.2018.06.001.

[8] J. Kolodner, *Case-Based Reasoning*. San Francisco, California, USA: Morgan Kaufmann Publishers Inc., 1993.

[9] M. M. Richter, R. O. Weber, *Case-Based Reasoning: A Textbook*. New York City, New York, USA: Springer Publishing Company, Incorporated, 2013.

[10] G. Costa Silva, E. E. O. Carvalho, W. M. Caminhas, "An artificial immune systems approach to Case-based Reasoning applied to fault detection and diagnosis," *Expert Systems with Applications*, vol. 140, p. 112906, Feb. 2020, doi: 10.1016/j.eswa.2019.112906.

[11] F. Torrent-Fontbona, J. Massana, B. López, "Case-base maintenance of a personalised and adaptive CBR bolus insulin recommender system for type 1 diabetes," *Expert Systems with Applications*, vol. 121, pp. 338–346, May 2019, doi: 10.1016/j.eswa.2018.12.036.

[12] E. Amador-Domínguez, E. Serrano, D. Manrique, J. F. D. Paz, "Prediction and decision-making in intelligent environments supported by knowledge graphs, A systematic review," *Sensors*, vol. 19, no. 8, p. 1774, 2019, doi: 10.3390/s19081774.

[13] "Ai4eu." https://www.ai4eu.eu/. Accessed: 2020-12-21.

[14] "The ai4eu scientific vision." https://www.ai4eu.eu/ai4eu-scientific-vision. Accessed: 2020-12-21.

[15] "Ai4eu." https://www.ai4eu.eu/resource/raidologist. Accessed: 2020-12-21.

[16] M. B. Bentaiba-Lagrid, L. Bouzar-Benlabiod, S. H. Rubin, T. Bouabana-Tebibel, M. R. Hanini, "A case-based reasoning system for supervised classification problems in the medical field," *Expert Systems with Applications*, vol. 150, p. 113335, July 2020, doi: 10.1016/j.eswa.2020.113335.

[17] D. Brown, A. Aldea, R. Harrison, C. Martin, I. Bayley, "Temporal case-based reasoning for type 1 diabetes mellitus bolus insulin decision support," *Artificial Intelligence in Medicine*, vol. 85, pp. 28–42, Apr. 2018, doi: 10.1016/j.artmed.2017.09.007.

[18] E. Lupiani, J. M. Juarez, J. Palma, R. Marin, "Monitoring elderly people at home with temporal Case-Based Reasoning," *Knowledge-Based Systems*, vol. 134, pp. 116–134, oct 2017, doi: 10.1016/j.knosys.2017.07.025.

[19] A. Aamodt, E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Commun.*, vol. 7, p. 39–59, Mar. 1994, doi: 10.3233/AIC-1994-7104.

[20] Y. Qin, W. Lu, Q. Qi, X. Liu, M. Huang, P. J. Scott, X. Jiang, "Towards an ontology-supported case-based reasoning approach for computer-aided tolerance specification," *Knowledge-Based Systems*, vol. 141, pp. 129–147, Feb. 2018, doi: 10.1016/j.knosys.2017.11.013.

[21] J. Daengdej, D. Lukose, R. Murison, "Using statistical models and case-based reasoning in claims prediction: experience from a real-world problem," *Knowledge-Based Systems*, vol. 12, pp. 239–245, Oct. 1999, doi: 10.1016/S0950-7051(99)00015-5.

[22] S. Nasiri, G. Zahedi, S. Kuntz, M. Fathi, "Knowledge representation and management based on an ontological CBR system for dementia caregiving," *Neurocomputing*, vol. 350, pp. 181–194, jul 2019, doi: 10.1016/j.neucom.2019.04.027.

[23] F. Marie, L. Corbat, Y. Chaussy, T. Delavelle, J. Henriet, J.-C. Lapayre, "Segmentation of deformed kidneys and nephroblastoma using Case-Based Reasoning and Convolutional Neural Network," *Expert Systems*

with Applications, vol. 127, pp. 282–294, Aug. 2019, doi: 10.1016/j.eswa.2019.03.010.

[24] L. Corbat, M. Nauval, J. Henriet, J.-C. Lapayre, "A fusion method based on Deep Learning and Case-Based Reasoning which improves the resulting medical image segmentations," *Expert Systems with Applications*, vol. 147, p. 113200, jun 2020, doi: 10.1016/j.eswa.2020.113200.

[25] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial Intelligence in Medicine*, vol. 94, pp. 42–53, Mar. 2019, doi: 10.1016/j.artmed.2019.01.001.

[26] V. Tang, K. Choy, G. Ho, H. Lam, Y. Tsang, "An iomt-based geriatric care management system for achieving smart health in nursing homes," *Industrial Management and Data Systems*, vol. 119, no. 8, pp. 1819–1840, 2019, doi: 10.1108/IMDS-01-2019-0024.

[27] S. Massie, G. Forbes, S. Craw, L. Fraser, G. Hamilton, "Fitsense: Employing multi-modal sensors in smart homes to predict falls," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11156 LNAI, pp. 249–263, 2018, doi: 10.1007/978-3-030-01081-2.

[28] G. Forbes, "Employing multi-modal sensors for personalised smart home health monitoring," vol. 2567, 2019, pp. 185–190.

[29] G. Mujtaba, L. Shuib, N. Idris, W. L. Hoo, R. G. Raj, K. Khowaja, K. Shaikh, H. F. Nweke, "Clinical text classification research trends: Systematic literature review and open issues," *Expert Systems with Applications*, vol. 116, pp. 494 – 520, 2019, doi: 10.1016/j.eswa.2018.09.034.

[30] J. T. Oliva, J. L. G. Rosa, "Classification for EEG report generation and epilepsy detection," *Neurocomputing*, vol. 335, pp. 81 – 95, 2019, doi: 10.1016/j.neucom.2019.01.053.

[31] S. Baccianella, A. Esuli, F. Sebastiani, "Variable-constraint classification and quantification of radiology reports under the ACR Index," *Expert Systems with Applications*, vol. 40, no. 9, pp. 3441 – 3449, 2013, doi: 10.1016/j.eswa.2012.12.052.

[32] C. R. Olsen, R. J. Mentz, K. J. Anstrom, D. Page, P. A. Patel, "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure," *American Heart Journal*, pp. 1–17, 2020, doi: 10.1016/j.ahj.2020.07.009.

[33] A. Dudchenko, M. Ganzinger, G. Kopanitsa, "Diagnoses Detection in Short Snippets of Narrative Medical Texts," *Procedia Computer Science*, vol. 156, pp. 150 – 157, 2019, doi: 10.1016/j.procs.2019.08.190.

[34] J. Prada, Y. Gala, A. Sierra, "Covid-19 mortality risk prediction using x-ray images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 7–14, 2021, doi: 10.9781/ijimai.2021.04.001.

[35] K. Negi, A. Pavuri, L. Patel, C. Jain, "A novel method for drug-adverse event extraction using machine learning," *Informatics in Medicine Unlocked*, vol. 17, p. 100190, 2019, doi: 10.1016/j.imu.2019.100190.

[36] "Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches," vol. 160, p. 113647, 2020, doi: 10.1016/j.eswa.2020.113647.

[37] T. F. d. Toledo, H. D. Lee, N. Spolaôr, C. S. R. Coy, F. C. Wu, "Web System Prototype based on speech recognition to construct medical reports in Brazilian Portuguese," *International Journal of Medical Informatics*, vol. 121, pp. 39 – 52, 2019, doi: 10.1016/j.ijmedinf.2018.10.010.

[38] L. F. Donnelly, R. Grzeszczuk, C. V. Guimaraes, W. Zhang, G. S. B. III, "Using a Natural Language Processing and Machine Learning Algorithm Program to Analyze InterRadiologist Report Style Variation and Compare Variation Between Radiologists When Using Highly Structured Versus More Free Text Reporting," *Current Problems in Diagnostic Radiology*, vol. 48, no. 6, pp. 524 – 530, 2019, doi: 10.1067/j.cpradiol.2018.09.005.

[39] H. Bay, T. Tuytelaars, L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006*, Berlin, Heidelberg, Germany, 2006, pp. 404–417, Springer Berlin Heidelberg.

[40] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, "Orb: An efficient alternative to sift or surf," in *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, USA, 2011, p. 2564–2571, IEEE Computer Society.

[41] P. F. Alcantarilla, A. Bartoli, A. J. Davison, "Kaze features," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ECCV'12, Berlin, Heidelberg, 2012, p. 214–227, Springer-Verlag.

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger Eds., Curran Associates, Inc., 2013, pp. 3111–3119.

[43] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[44] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, "From word embeddings to document distances," vol. 37 of *Proceedings of Machine Learning Research*, Lille, France, 07–09 Jul 2015, pp. 957–966, PMLR.

[45] W. Boag, E. Sergeeva, S. Kulshreshtha, P. Szolovits, A. Rumshisky, T. Naumann, "Cliner 2.0: Accessible and accurate clinical concept extraction," in *ML4H: Machine Learning for Health Workshop at Advances in Neural Information Processing Systems*, NIPS '17, 2017.

[46] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, 09 2019, doi: 10.1093/bioinformatics/btz682.

[47] I. Beltagy, K. Lo, A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3615–3620, Association for Computational Linguistics.

[48] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, p. 317, Dec 2019, doi: 10.1038/s41597-019-0322-0.

[49] D. Demner-Fushman, S. Antani, M. Simpson, G. R. Thoma, "Design and development of a multimodal biomedical information retrieval system," *Journal of Computing Science and Engineering*, vol. 6, no. 2, pp. 168–177, 2012, doi: 10.5626/JCSE.2012.6.2.168.

[50] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, Feb. 1966.

Elvira Amador-Domínguez

Elvira Amador-Domínguez received the B.Sc. degree in Computer Science (2017) and the M.Sc. in Artificial Intelligence (2018) from the Universidad Politécnica de Madrid. Her B.Sc final thesis was awarded as one of the best thesis of the year 2017. She is currently a PhD Candidate at the Universidad Politécnica de Madrid, founded by a grant of the own university. Her prime fields of research include knowledge representation, deep learning, knowledge integration and explainability. She has also participated in the European Project AI4EU, as well as in a national educational innovation project.

Emilio Serrano

Emilio Serrano received the M.Sc. degree in computer science (2006) and the Ph.D. degree, with European mention and Extraordinary Ph.D. Award in artificial intelligence (2011), from the University of Murcia, Spain. He has also been a Visiting Researcher with The University of Edinburgh, the University of Oxford, and the National Institute of Informatics in Tokyo. He is currently an Associate Professor with the Department of Artificial Intelligence, Universidad Politécnica de Madrid (UPM). He is also Secretary of the Ph.D. in Artificial Intelligence at UPM. His main research line is the Social and Explainable Artificial Intelligence for Smart Cities. His scientific production includes more than 80 publications, highlighting over 25 articles in the JCR. He lectures deep learning and social network analysis among other courses; and, has been principal investigator in three educational innovation projects in data science. He has also participated in several European and National funding programs such as FP7 research projects (smartopendata, eurosentiment, and omelette) and H2020 research projects (slidewiki and AI4EU).

Daniel Manrique

Daniel Manrique received the B.S. and Ph.D. degrees in computer science from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 1997 and 2001, respectively. He has been a visiting researcher with the University of Sunderland and Trinity College of Dublin. He is currently a member of the artificial intel ligence lab workgroup and an Associate Professor of computing with the Departmento de Inteligencia Artificial, UPM's School of Computing. His major fields of study and research are the subsymbolic artificial intelligence, its synergies with the symbolic domain, and diverse applications such as the medical area. He has published more than 70 research works on these topics in international journals, conferences, books, and book chapters. Dr. Manrique has participated as a researcher in several European, National, and regional research projects related. He is a member of the international program committee of several international congresses and acts as a reviewer of impact journals in the Journal Citation Report.

Javier Bajo

Dr. Javier Bajo, full professor at the Department of Artificial Intelligence, Computer Science School at Universidad Politécnica de Madrid (UPM), holds (since 03/05/2019) the position of Director of the UPM AI.nnovation Space Research Center in Artificial Intelligence. He was Director of the Department of Artificial Intelligence (20/05/2016-19/10/2017) at UPM, Secretary of the PhD in Artificial Intelligence at UPM (23/06/2016-19/10/2017) and Coordinator of the Research Master in Artificial Intelligence at UPM (18/02/2013 - 20/05/2016). He also holds the position of Director of the Data Center at the Pontifical University of Salamanca (13-10/2010 - 08-11-2012), with 21 employees. His main lines of research are Social Computing and Artificial and Hybrid Societies; Intelligent Agents and Multiagent Systems, Ambient Intelligence, Machine Learning. He has supervised 11 Ph.D thesis, participated in more than 50 research projects (in most of them as principal investigator) and published more than 300 articles in recognized journals (81 JCR papers) and conferences. His h-index is 39. He is founder of the PAAMS series of conferences and is an IEEE, ACM and ISIF member.

Using Grip Strength as a Cardiovascular Risk Indicator Based on Hybrid Algorithms

E. F. Bareño-Castellanos[1]*, P. A. Gaona-García[1], J. E. Ortiz-Guzmán[2], C. E. Montenegro-Marin[1]

[1] Faculty of Engineering, Universidad Distrital Francisco José de Caldas, Bogotá (Colombia)
[2] Facultad de Ciencias de la Salud, Universidad de Ciencias Aplicadas y Ambientales, Bogotá (Colombia)

Received 6 November 2020 | Accepted 9 March 2021 | Published 6 May 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

This article shows the application and design of a hybrid algorithm capable of classifying people into risk groups using data such as prehensile strength, body mass index and percentage of fat. The implementation was done on Python and proposes a tool to help make medical decisions regarding the cardiovascular health of patients. The data were taken in a systematic way, k-means and c-means algorithms were used for the classification of the data, for the prediction of new data two vectorial support machines were used, one for the k-means and the other for the c-means, obtaining as a result a 100% of precision in the vectorial support machine with c-means and a 92% in the one of k-means.

Keywords

Body Mass Index, C-Means, K-Means, Percentage F Fat, Prehensile Strength, Risk Indicator, Support Vector Machine.

DOI: 10.9781/ijimai.2021.05.004

I. Introduction

CARDIOVASCULAR diseases are the leading cause of mortality in the world, with ischemic heart disease and cerebrovascular disease being the most frequent [1]. There are factors such as obesity, sedentary lifestyle, hypertriglyceridemia, smoking and family history that are directly related to heart problems and are used by specialists for the early detection of heart disease [2]. For this reason, the importance of finding mechanisms for the assessment of risk factors associated with cardiovascular diseases is growing.

Currently studies are presented where grip strength has received considerable attention not only for the relationships of protein level, muscle mass and strength, but as an indicator of physical capacity and general health [3]. Some studies have suggested that grip strength is a risk factor in multiple diseases such as diabetes, cancer and different heart diseases [4], therefore it is considered that a strong or weak grip of the hand carries more than social cues, as it can also help measure an individual's risk of having a heart attack or stroke, or dying from cardiovascular disease [5].

One of the areas that could contribute in the assessment of these indicators is through processes based on data mining, given its relevance to obtain knowledge of large amounts of data. The data mining-based process enables the integration of multiple disciplines such as statistics, machine learning, neural networks, and pattern recognition. Authors such as [6]-[7] have used techniques such as the nearest neighbor (KNN), decision trees (DT), genetic algorithms (GA) and naive Bayesian classifier (NB) to early detection of heart problems.

Based on the above, a model based on Support Vector Machine (SVM) is presented for classification in risk groups using clustering techniques that allow establishing relationships between grip strength and different physiological variables such as height, weight, body mass index and fat percentage. One of the purposes of this study is to evaluate the behavior and reliability of algorithms based on SVM, so that later a tool can be proposed that can provide support to an expert to determine health status in a quick and complementary way.

For the grouping algorithms two models were used, one of them is a c-means hybrid model and the second one is k-means, to feed these models a dataset was used with data such as age, height, weight, fat percentage and caporal mass index. In addition, the prehensile force was used as a predictive variable in the vectorial support machine.

The following article is structured as follows: section II, related works presents all the background of the research, as well as related works. Section III establishes the work methodology to be followed. Section IV proposes the data mining model for the case study, section V shows the results obtained. Finally, section VI is the discussion of results and conclusions are presented in section VII.

II. Related Works

Biomarkers are medical indicators that demonstrate bodily functions or pathological processes in addition to providing an objective indication of medical status [8]. The use of grip strength as a biomarker has taken on great importance. In the latest research, it has been determined that low levels of grip strength are associated with various associated morbidities; an increased risk of falls, hospital stay and mortality; and a lower quality of life [9]. For example, the Prospective Urban-Rural Epidemiology (PURE) study determined that the decrease in grip strength is a risk factor for the incidence of

* Corresponding author.

E-mail address: efbarenoc@correo.udistrital.edu.co

cardiovascular diseases and can predict the risk of death in people who develop cardiovascular or non-cardiovascular diseases [10].

Within the context of analysis, data mining is the process of extracting information in order to obtain knowledge from large amounts of data. It is an integration of multiple disciplines such as statistics, machine learning, and neural networks [11]. Data mining allows the exploration of large data sets in search of relationships, knowledge and patterns that are difficult to determine with traditional methods.

In recent years, different machine learning techniques have been used to predict heart disease, using a database with variables such as blood pressure, cholesterol, blood sugar, weight, age and sex. Some of these techniques will be discussed below. Authors such as Rairikar et al. used three data mining techniques to predict heart diseases, which were decision tree, random forests and Naive Bayes, determined that the random forest is the algorithm with the highest precision in addition to the use of a genetic algorithm to select the most important characteristics of the data sets [6]. In a study carried out by Bahrami & Hosseini Shirvani, they verified the different classification techniques in the diagnosis of heart diseases, using techniques such as the decision tree and KNN, the Naive Bayes algorithm was also used for the grouping of relevant characteristics. After ranking and performance evaluation, the decision tree is considered the best algorithm for heart disease diagnosis with the data set that was chosen [7].

For their part Jabbar et al. proposed a new hybrid algorithm that combines the KNN algorithm with a genetic algorithm, they used the genetic algorithm to classify the most relevant attributes of the dataset and then used the KNN algorithm, the experimental results showed that the use of a genetic algorithm improved accuracy in diagnosing heart disease [12]. Authors such as Sowmiya used different machine learning techniques in multiple datasets to determine which are the best techniques in the use of prediction of heart disease, diseases such as fibrillation, congenital heart disease, coronary artery disease, heart attack were studied [13]. Along the same lines, Gawande & Barhatte used a 7-layer convolutional neural network for the classification of heart diseases through the analysis of ECG signals (electrocardiogram), achieving a precision of 99.46% in different patients [14]. For their part, Amin et al. carried out a study where they would evaluate multiple algorithms in the prediction of heart problems, the data set used contained the variables: age, sex, Cp, pressure in each person, amount of sugar in the blood, ECG results, maximum heart rate, and heart status. It was determined that the algorithm with the best accuracy was a Support Vector Machine with a result of 86.87% with the use of 9 attributes [15].

The authors López-Martínez et al used a neural network model to predict neonatal sepsis, using a data set of 555 patients divided between 66% positive and 34% negative cases, resulting in an accuracy of 83.1% [16]. In the same way the authors López-Martínez et al. used a neuronal network to estimate the association of variables such as BMI, race, age, among others with hypertension reaching a specificity of 87% with a precision of 57.8% [17].

A work that can be highlighted is the one of Devi et al., which uses the diffuse c-means model for the grouping of skin lesions using non-dermoscopic images resulting in an accuracy of 95.69% and a sensitivity of 90.02% [18]. The problem of this work is that it is only for non-dermoscopic images. The next work by López-Martínez, Núñez-Valdez, García-Díaz, et al., uses big data and artificial intelligence to improve the support to medical decisions in health management of the population, this work resulted in the detection of patterns using different types of data [19].

Taking into account the aforementioned works, the following study aims to use grip strength as a risk indicator for heart problems, data mining algorithms (machine learning) will be used to determine relationships with other biomarkers such as weight, height, sex, percentage of fat and BMI among others. The objective of this work is to create a tool that can support specialists in decision-making in the medical field and health in general.

III. Methodology

In this study, machine learning methods were used to create a model based on Support Vector Machine and clustering, this was trained with a database obtained systematically at the university of Rosario. Fig. 1 shows the process of obtaining, cleaning, analysis and results of this investigation.

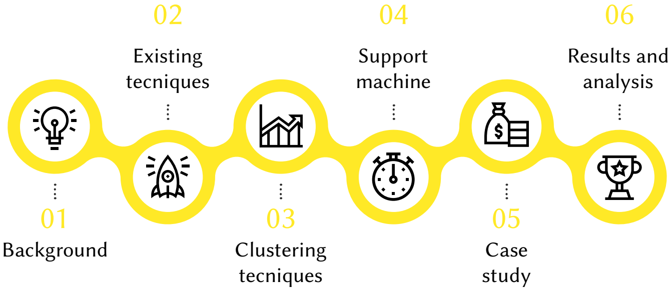


Fig. 1. Methodology and steps for its application.

A. Sample Population

The sample included in this research corresponded to students from the School of Medicine and Health Sciences of the Universidad del Rosario (Bogotá, Colombia), who were evaluated in the research laboratory of the Physiology Unit. In total, 80 students (50 women and 30 men) were included in the study. The characteristics of the Sample are reported in Table I.

TABLE I. Description of the Sample Composed of 29 Men and 50 Women, the Table Contains the Average (A), the Deviation (D), the Minimum (MI) and the Maximum (MA). The Age Is in Years, the Height Is in Centimeters, the Weight Is in Kilograms, the Body Mass index (BMI) Is in Kilograms Per Square Meter and the Maximum Grip force Is in Newtons

	Men n=29				Women n=50			
	A	D	MI	MA	A	D	MI	MA
Age	18,1	1,29	17	22	17,83	1,04	16	20
Height	175,41	5,56	163	186	161,14	6,83	145	177
Weight	70,19	11,17	53,3	95,1	58,29	8,28	48,2	76,4
% fat	16,76	6,74	6	30,4	31,94	4,87	20,6	41,7
BMI		3,54	16,5	31,8	22,38	2,35	17,2	28,4
grip force		22,16	21,35	108,86	29,1	9,36	17,12	65,76

The dataset can be found through github[1].

The sampling that was followed in this study corresponded to a convenience sampling, since the participants were included after verifying that they met the inclusion criteria and that they did not have any exclusion criteria that did not allow them to be included in the study sample. Regarding the above, the inclusion criteria were: being of legal age, accepting participation in this study by signing the informed consent, not having consumed caffeinated beverages in the last 2 hours or having practiced any intense physical activity moderate or high on the last day and complete all the protocols required to record Heart Rate Variability (HRV) and grip strength. Regarding the exclusion criteria, those participants who were consuming any medication with direct action on the nervous system or on cardiovascular function, who had consumed beverages rich in

[1] https://github.com/FrederickUdis/data-Grip-strength.git

caffeine or stimulants of the central nervous system in the last 2 hours were not taken into account. Besides, those who had not signed the informed consent and who did not complete all the protocols required for data collection were excluded.

B. Methodology

HRV Variability was recorded through Polar V800TM reference monitors, which have been widely validated in the scientific literature for HRV recording. The records were taken in a sitting position, for 15 minutes, in a room at stable temperature and relative humidity. During this period, all participants remained completely silent, without control of the ventilatory rate and trying to be as relaxed as possible. The Polar record was digitized using the Polar ProTrainerTM program and HRV analysis was performed in the KubiosTM program with a 5-minute window of the total record, from which variables were extracted.

On the other hand, the grip force was recorded using the PowerLabTM system, using the manual dynamometer from the same developer. In total, each participant made 3 attempts, always with the dominant hand and the arm flexed at 90 ° with the elbow as close as possible to the participant's trunk and without moving them neither forward nor backward during the recording. The procedure consisted of squeezing the dynamometer as hard as possible for 3 seconds and this action was repeated 3 times, with a 5-minute rest period between each series. Of the 3 attempts, the best was chosen for the final analysis.

The second part of the first phase consisted of a background study where a rigorous investigation was carried out on the grip force as an indicator of heart problems, as well as the different machine learning techniques used in heart problems. In the second phase, the different techniques sought in phase one were evaluated, determining that the most appropriate algorithms for our problems are two types of algorithms, the first group is determined supervised algorithms and the other is unsupervised.

In the third phase, an unsupervised algorithm was used because [20] it does not depend on specific instructions when performing a classification, instead it is based on the autonomous grouping of data through exploration, this type of learning is very similar to the way humans learn, which offers a very flexible approach when applying it to our problem. It should be noted that in addition to an unsupervised algorithm, a hybrid algorithm was used that uses fuzzy logic rules to increase precision.

In phase four the model based on Support Vector Machine was used to classify the results, this was applied through a case study in phase five.

The result and analysis are carried out in phase six, in addition to obtaining the performance metrics of the algorithm.

IV. Case Study

The proposed model has two main components that can be observed in Fig. 2, the first is the data processing and the second is the data analysis, in the data processing during the data acquisition we obtain the data in raw and we go to the preparation of the data where we do a cleaning to convert it into a dataset for analysis, the next component has three parts (analysis, results and conclusions).

The main part of the second component is the analysis, which is divided into 4 subcomponents that are the data preprocessing, the clusters, the classification model and finally the output, in section 4.1 to 4.4 we will talk more about these components.

Based on the above, the model is a 2-layer model, the first layer being in charge of acquiring and preparing the data for the second layer, this is done with the Python programming language and the specialized libraries for data treatment.

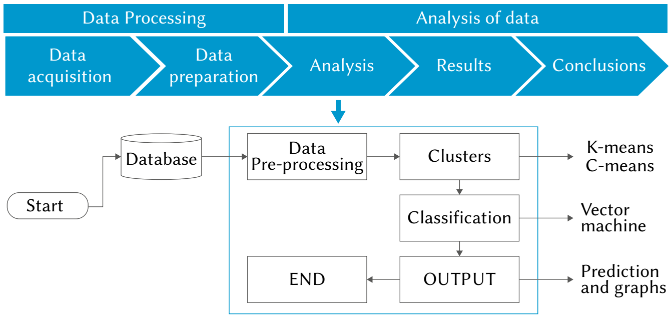


Fig. 2. Model proposed for data processing and data analysis.

The second layer is the constructed algorithm that has two important parts that are the grouping where the two chosen algorithms are used to later pass to the Support Vector Machine which has the function of classifying the result of the grouping.

A. Data Cleaning

After data collection. All records with missing values were removed from the data set, reducing the number of data from 80 to 78. Then the variable sex (male or female) became a category variable or more commonly called binary (0 or 1). The task of data preprocessing was carried out by cleaning null data, the data was divided between men and women. Subsequently, the characteristics of the names and surnames were eliminated to give a total of 19 characteristics.

B. Data Processing

After processing the data, two clustering algorithms were used to determine the risk groups, [21] describes the k-means clustering algorithm as a popular algorithm due to its effectiveness in dividing multiple points into k clusters, likewise it has a high adaptability to different problems. For our study, this was the first of the two clustering algorithms that were used, the purpose of this clustering was to find groups of related data establishing relationships between characteristics such as grip strength (grip strength) and body mass index (BMI), fat percentage (% Fat) among others. It works by adding a new column to the dataset which contains the labels of the cluster to which each data belongs to finally be classified by a Support Vector Machine (SVM).

The second algorithm used was the c-means (FMC) which assigns a probability of belonging to each point on a group or cluster to finally choose the best group that belongs to the said point [22].

The implementation of the k-means algorithm was done with the Kmeans library of sklearn, three groups or clusters which are determined by 3 stars were obtained as a result. This can be seen in Fig. 3.

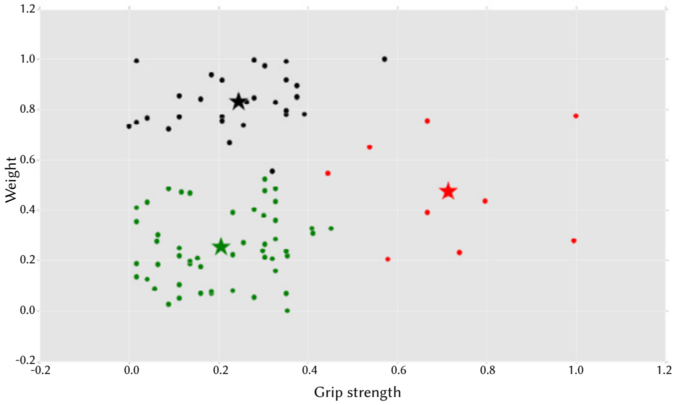


Fig. 3. Result obtained from the k-means algorithm using the python language for analysis and visualization.

The Liberia scikit-fuzzy, which contains the skfuzzy package, was used for the application of the c-means algorithm. An evaluation of up to a maximum of 10 clusters determined that the optimal number of clusters is 2, this can be seen in Fig. 4.

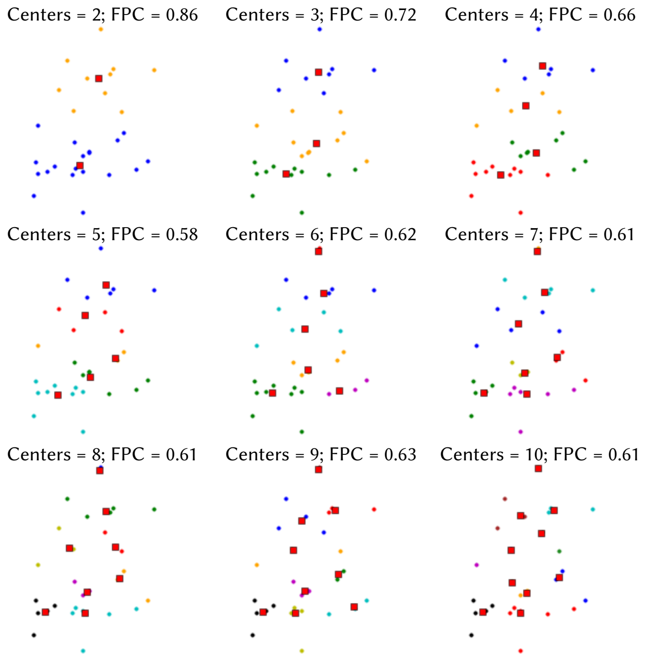
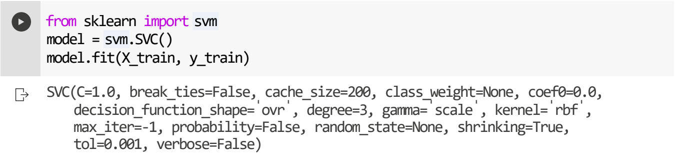


Fig. 4. Result of the c-means algorithm from 2 to 10 clusters.

C. Classification

Pitale et al. [23] establishes two steps for the application of the classification algorithms: the first phase consists of the definition of the model, in the second phase the method is selected and finally a method is applied to classify it. For this study, a Support Vector Machine (SVM) was used, which aims to classify new data in three risk regions (low-medium-high) according to grip strength and fat percentage.

The vectorial support machine was made in Python with the sklearn. SVM library which offers everything necessary to apply the vectorial support machine as shown in Fig. 5.



```
from sklearn import svm
model = svm.SVC()
model.fit(X_train, y_train)
```

```
SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
```

Fig. 5. Code for the use of the sklearn library in addition to the application of the Support Vector Machine (SVM).

V. Results and Analysis

Of the 79 participants in the analysis, 29 were men and 50 were women. 48.27% of the men had percentage fat greater than 20% and a BMI greater than 20 kg / m2 in addition to grip strength of 40 and 70 Newtons.

As for the women, 78% had a high percentage of fat above 30% but only 14% had a BMI above 25 kg/m 2 in addition to a grip strength of 24 to 40 as shown in Tables II, III, IV. The description in Table II provides more data on the results obtained. There was no significant difference in terms of age, height and weight.

TABLE II. Body Mass index (BMI) Measurements, the Following Parameters Were Used to Obtain the Measurements Corresponding to Table I: Insufficient Weight <18,5 Kg/m2, Normal Weight 18,5 - 24,9 Kg/m2, Overweight Degree I 25,0-26,9 Kg/m2, Overweight Degree II 27,0-29,9 Kg/m2, Type I obesity 30,0-34,9 Kg/m2, Type II Obesity 35,0-39,9 Kg/m2, Type III Obesity 40,0-49,9 Kg/m2 and Type IV Obesity 40,0-49,9 Kg/m2, these Parameters Were Obtained from [25]

	Men n = 29		Women n = 50	
	F	%	F	%
Insufficient	1	3,45	4	8
Normal	21	72,41	39	78
Overweight	5	17,24	4	8
Overweight	0	0	2	4
Type I obesity	2	6,9	1	2
Type II obesity	0	0	0	0
Type III obesity	0	0	0	0
Type IV obesity	0	0	0	0

TABLE III. Guidelines for Determining which Subjects Had a Large Percentage of Body Fat

	Fat% limit values	
characteristics	MEN	WOMEN
thin	<8%	<15%
Optimum	8,1 a 15,9%	15,1 a 20,9%
Slightly overweight	16,0 a 20,9%	21,0 a 25,9%
Overweight	21,0 a 24,9%	26,0 a 31,9%
Obesity	≥25%	≥32%

TABLE IV. Measurements of Fat Percentage, these Values Were Obtained Through the Grouping Analysis Carried Out Before the Prediction, the Results Obtained Determined that Women Possess a higher Percentage of Fat than Men, as Determined in [25].

Men n = 29		Women n = 50	
F	%	F	%
4	13,79	0	0
11	37,93	4	8
5	17,24	4	8
6	20,69	19	38
3	10,34	23	46

According to Table IV, it was possible to establish a group of women who have a percentage Fat greater than 32% with a grip force that ranges between 20 and 40 Newtons, this can be seen in Fig. 6, which implies that women with obesity have a grip strength of less than 30, which is the average seen in Table I.

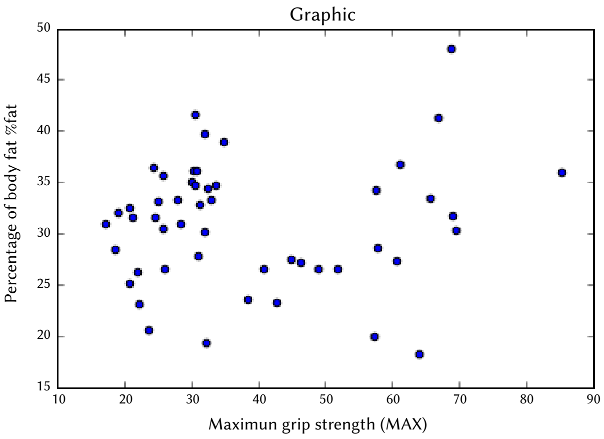


Fig. 6. Dispersion graph of the data referring to the women of the percentage of fat (% fat) Vs maximum grip strength (MAX), performed with Python.

Based on table IV, 90% of this group also has a BMI greater than 22 kg / m2, which indicates a clear relationship between overweight and grip strength in women with these characteristics; this can be clearly seen in Fig. 7.

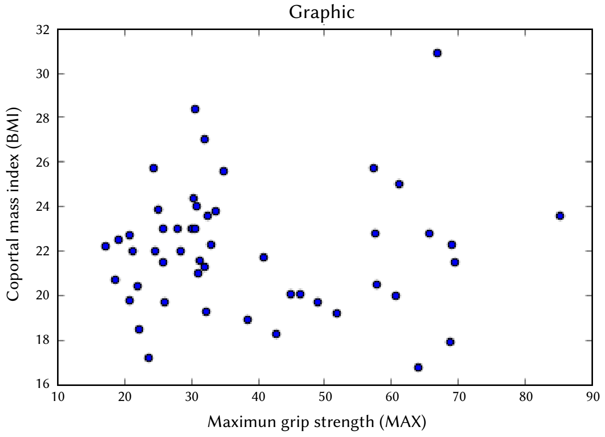


Fig. 7. Dispersion graph of the data referring to the body mass index (BMI) Vs maximum grip strength (MAX), performed with Python.

With respect to men, the group cannot be clearly seen due to the lack of data, but data that can be used as a group or clusters can be seen, this can be seen in Fig. 8 and Fig. 9.

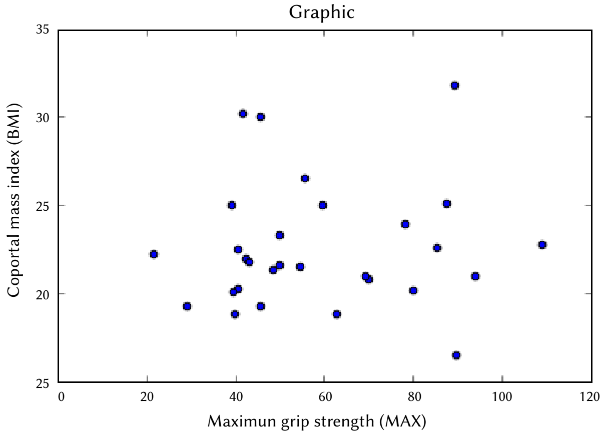


Fig. 8. Dispersion graph of the data referring to the men's body mass index (BMI) Vs maximum grip strength (MAX), performed with Python.

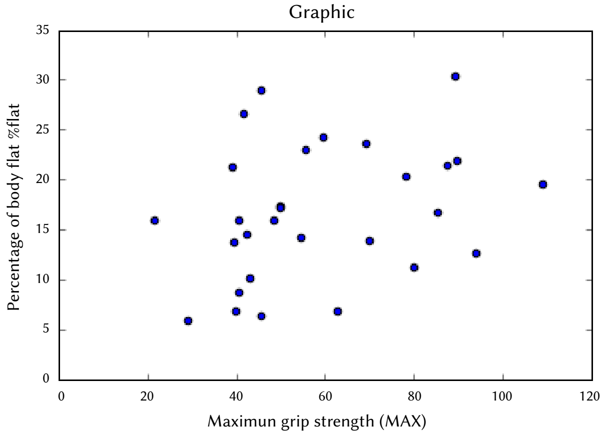


Fig. 9. Dispersion graph of the data referring to men of the percentage of fat (% Fat) (BMI) Vs maximum grip strength (MAX), performed with Python.

According to the results of the use of the k-means technique, 3 risk groups were revealed (low, medium, high). A Support Vector Machine was used to classify these groups and predict new data in each of these groups. Fig. 10 shows the classification of the Support Vector Machine.

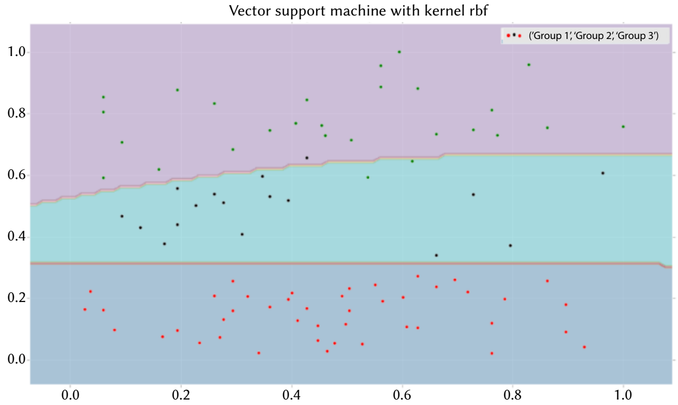


Fig. 10. Result of Support Vector Machine training for data classification using python and sklearn.

In other part, the results of the hybrid C-means technique, determined the existence of two risk groups, with the help of the Support Vector Machine it was possible to classify new data into these two groups. Fig. 11 shows the results of the Support Vector Machine applied to the results of the c-means algorithm.

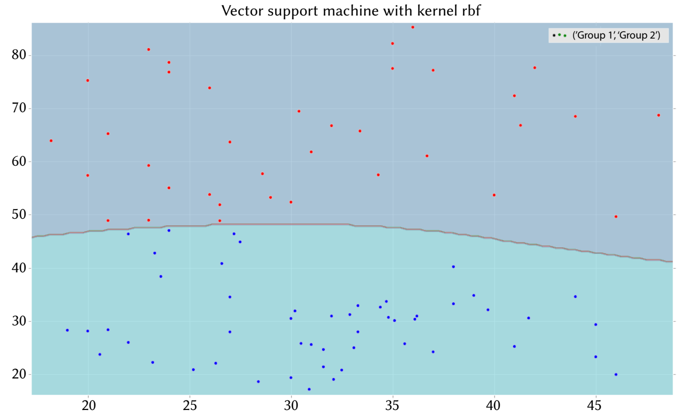
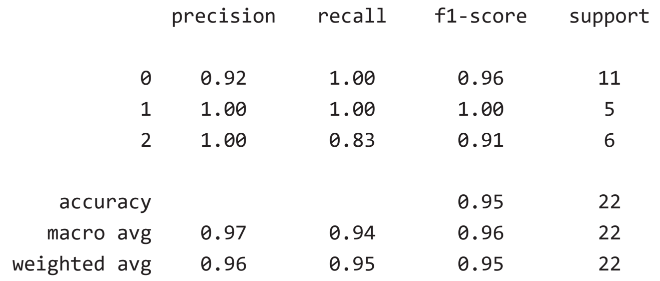


Fig. 11. Result of Support Vector Machine training for data classification using python and sklearn.

To identify the best combination of algorithms, the confusion matrix incorporated in the Python sklearn library was used. This matrix evaluates multiple metrics to determine the best algorithm. The evaluation parameters used are:

- Accuracy
- Recall
- F1-score
- Support

These parameters are obtained through the confusion matrix, which is an evaluation tool used in the machine learning area. The columns of a Confusion Matrix represent the results of the prediction class, and the rows represent the results of the class [24]. Fig. 12 shows the results of the Support Vector Machine with the k-means algorithm.

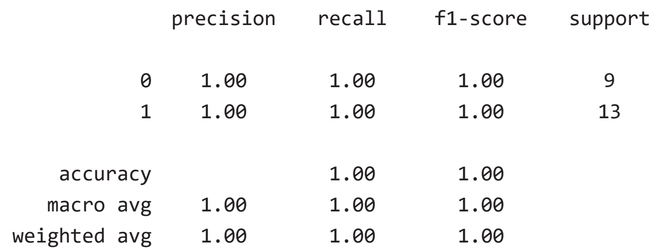


```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        11
           1       1.00      1.00      1.00         5
           2       1.00      0.83      0.91         6

    accuracy                           0.95        22
   macro avg       0.97      0.94      0.96        22
weighted avg       0.96      0.95      0.95        22
```

Fig. 12. Metrics of the k-means algorithm and Support Vector Machine obtained through the sklearn python library.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         9
           1       1.00      1.00      1.00        13

    accuracy                 1.00      1.00
   macro avg       1.00      1.00      1.00
weighted avg       1.00      1.00      1.00
```

Fig. 13. Metrics of the c-means algorithm and Support Vector Machine obtained through the sklearn python library.

According to the results obtained, they can be considered acceptable, since the two algorithms have an accuracy higher than 90%, being the best the c-means algorithm with the support vector machine, which presents an accuracy of 100%, the high accuracy rate of the two algorithms is due to the low amounts of data obtained, which are not more than 80. By reducing the number of variables it was possible to obtain great precision with the two algorithms and to determine the influence of the percentage of fat in the prehensile force in women, this was achieved thanks to the use of unsupervised algorithms (c-means and k-means), the use of vectorial support machines gave good results and can be used as a tool to help heart specialists.

VI. Discussion

In this study it was possible to demonstrate that male volunteers who had a low percentage body fat (<23%) and a lower BMI had a grip strength level of less than 40 Newtons, compared to people who have a body mass index and level of average fat, aspect that corroborates the work carried out by Alberto Cardozo [25]. However, in women it was presented that body weight and BMI were higher than in men, coinciding with what was found by other authors [26], although the latter presented a higher percentage of fat (> 30%); These elements could be induced by various factors such as hormones, eating habits, body composition, etc. Regarding the grip strength, it was found that several women who had percentage of fat greater than 28% had a grip strength less than 40 Newtons, but the BMI was normal.

During the creation and evaluation of the algorithm, multiple problems and drawbacks were found, such as the lack of important characteristics associated with smoking and comorbidities in the dataset. Similarly, at the time of making the correlation matrix, relationship values of less than 30% were found, for this reason it was decided to perform unsupervised algorithms to validate relationships that are not seen in the correlation matrix. In addition, the model presented a problem of a perfect fit to the training set, this situation was caused by the fact that the amount of training data was less compared to the greater number of characteristics, a very common problem presented by supervised classification algorithms [27].

Finally, to highlight, the average value of precision of the applied algorithms were higher in precision than the results obtained by (Amin et al., 2019), since when they used a Support Vector Machine fed with 9 variables, they obtained a precision of 86% while we obtained 100% accuracy due to the use of a clustering algorithm before using the Support Vector Machine. On the contrary, the work carried out by Gawande & Barhatte used a 7-layer neural network using electrocardiograms to find heart problems and obtained an accuracy of 98.7 [14].

VII. Conclusion

With the use of clustering techniques (k-means and c-means), groups with low grip strength relationships and high fat percentage were found, as well as the body mass index, which indicates a relationship. Likewise, the use of a Support Vector Machine allowed the classification of new data into risk groups.

One of the main advantages of using easily obtained variables with grouping and prediction algorithms was the rapid classification of new data, which can be used as a tool for medical decision-making in patients with cardiovascular risk. These algorithms can be applied to any type of patient consulted by a specialist, as long as more variables that imply cardiovascular risk are taken into account.

It is recommended for use in conjunction with tools capable of detecting heart problems and under the supervision of qualified specialists.

The proposed algorithms are suggested as a useful and complementary tool for people who want to know their risk group, since it combines variables that can be easily obtained such as grip strength, weight and body mass index.

As a future study, it is expected to complement the experiment with additional variables associated with smoking, sedentary lifestyle and family history of cardiac comorbidities associated with grip strength. Likewise, other classification techniques such as neural networks, decision trees or genetic algorithms, will be evaluated.

References

[1] H. Wang et al., “Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015,” The Lancet, vol. 388, no. 10053, pp. 1459–1544, 2016, doi: 10.1016/S0140-6736(16)31012-1.

[2] D. Mercy et al., “Riesgo cardiovascular global y edad vascular: herramientas claves en la prevención de enfermedades cardiovasculares Global cardiovascular risk and vascular age: key tools in cardiovascular diseases prevention,” Revista Médica Electrónica, Artic. Revisión, pp. 211–226, 2014.

[3] O. Prasitsiriphon and W. Pothisiri, “Associations of Grip Strength and Change in Grip Strength With All-Cause and Cardiovascular Mortality in a European Older Population,” Clinical Medicine Insights: Cardiology, vol. 12, 2018, doi: 10.1177/1179546818771894.

[4] C. A. Celis-Morales et al., “Associations of grip strength with cardiovascular, respiratory, and cancer outcomes and all cause mortality: Prospective cohort study of half a million UK Biobank participants,” BMJ (Online), vol. 361, pp. 1–10, 2018, doi: 10.1136/bmj.k1651.

[5] M. D. Howard LeWine, “Grip strength may provide clues to heart health - Harvard Health Blog - Harvard Health Publishing,” Grip strength may provide clues to heart health, 19-May-2015. [Online]. Available: https://www.health.harvard.edu/blog/grip-strength-may-provide-clues-to-heart-health-201505198022. [Accessed: 03-Jul-2020].

[6] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, “Heart disease prediction using data mining techniques,” Proceedings of 2017 International Conference on Intelligent Computing and Control, I2C2 2017, vol. 2018-Janua, no. October, pp. 1–8, 2018, doi: 10.1109/I2C2.2017.8321771.

[7] B. Bahrami and M. Hosseini Shirvani, “Prediction and Diagnosis of

Heart Disease by Data Mining Techniques," Journal of Multidisciplinary Engineering Science and Technology, vol. 2, no. 2, pp. 3159–40, 2015.

[8] R. W. Bohannon, "Grip strength: An indispensable biomarker for older adults," Clinical Interventions in Aging, vol. 14, pp. 1681–1691, 2019, doi: 10.2147/CIA.S194543.

[9] sevtap güllüoğlu badıl, "Does Vitamin D Level Affect Grip Strength: a Cross-sectional Descriptive Study," Erciyes Medical Journal, vol. 42, no. 1, pp. 7–11, 2019, doi: 10.14744/etd.2019.15428.

[10] D. P. Leong et al., "Prognostic value of grip strength: Findings from the Prospective Urban Rural Epidemiology (PURE) study," The Lancet, vol. 386, no. 9990, pp. 266–273, 2015, doi: 10.1016/S0140-6736(14)62000-6.

[11] B. Deekshatulu and P. Chandra, "Classification of heart disease using artificial neural network and feature subset selection," Global Journal of Computer Science and Technology, vol. 13, no. 3, 2013.

[12] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technology, vol. 10, pp. 85–94, 2013, doi: 10.1016/j.protcy.2013.12.340.

[13] C. Sowmiya, "Comparative Study of Predicting Heart Disease By Means Of Data Mining," Int. J. Eng. Comput. Sci., vol. 5, no. 12, pp. 19580–19582, 2016.

[14] N. Gawande and A. Barhatte, "Heart diseases classification using convolutional neural network," Proceedings of the 2nd International Conference on Communication and Electronics Systems, ICCES 2017, vol. 2018-Janua, no. Icces, pp. 17–20, 2018, doi: 10.1109/CESYS.2017.8321264.

[15] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, vol. 36, pp. 82–93, 2019, doi: 10.1016/j.tele.2018.11.007.

[16] F. López-Martínez, E. R. Núñez-Valdez, J. Lorduy Gomez, and V. García-Díaz, "A neural network approach to predict early neonatal sepsis," Computers and Electrical Engineering, vol. 76, pp. 379–388, 2019, doi: 10.1016/j.compeleceng.2019.04.015.

[17] F. López-Martínez, E. R. Núñez-Valdez, R. G. Crespo, and V. García-Díaz, "An artificial neural network approach for predicting hypertension using NHANES data," Scientific Reports, vol. 10, no. 1, pp. 1–14, 2020, doi: 10.1038/s41598-020-67640-z.

[18] S. S. Devi, N. H. Singh, and R. H. Laskar, "Fuzzy C-Means Clustering with Histogram based Cluster Selection for Skin Lesion Segmentation using Non-Dermoscopic Images," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, p. 26, 2020, doi: 10.9781/ijimai.2020.01.001.

[19] F. López-Martínez, E. R. Núñez-Valdez, V. García-Díaz, and Z. Bursac, "A case study for a big data and machine learning platform to improve medical decision support in population health management," Algorithms, vol. 13, no. 4, pp. 1–19, 2020, doi: 10.3390/A13040102.

[20] A. Roohi, K. Faust, U. Djuric, and P. Diamandis, "Unsupervised Machine Learning in Pathology: The Next Frontier," Surgical Pathology Clinics, vol. 13, no. 2, pp. 349–358, 2020, doi: 10.1016/j.path.2020.01.002.

[21] S. K. Majhi and S. Biswal, "Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer," Karbala International Journal of Modern Science, vol. 4, no. 4, pp. 347–360, 2018, doi: 10.1016/j.kijoms.2018.09.001.

[22] V. Manikandan, V. Porkodi, A. S. Mohammed, and M. Sivaram, "Ensemble Classification Based Microarray Gene Retrieval System," ICTACT Journal on Soft Computing, vol. 9, no. 1, pp. 1806–1812, 2018, doi: 10.21917/ijsc.2018.0252.

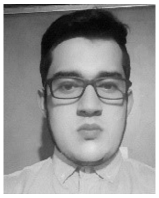
[23] R. Pitale, K. Tajane, and J. Umale, "Heart Rate Variability Classification and Feature Extraction Using Support Vector Machine and PCA: An Overview," Journal of Engineering Research and Applications www.ijera.com, vol. 4, no. 1, pp. 381–384, 2014.

[24] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Information Sciences, vol. 507, pp. 772–794, 2020, doi: 10.1016/j.ins.2019.06.064.

[25] L. Alberto Cardozo, L. Alberto, C. Guzman, Y. Andrés, M. Torres, and J. Alejandro, "Artículo Original Porcentaje de grasa corporal y prevalencia de sobrepeso-obesidad en estudiantes universitarios de rendimiento deportivo de Bogotá, Colombia Body fat percentage and prevalence of overweight-obesity in college students of sports performanc," Nutrición clínica y dietética hospitalaria, vol. 36, no. 3, pp. 68–75, 2017, doi: 10.12873/363cardozo.

[26] M. E. Piché, P. Poirier, I. Lemieux, and J. P. Després, "Overview of Epidemiology and Contribution of Obesity and Body Fat Distribution to Cardiovascular Disease: An Update," Progress in Cardiovascular Diseases, vol. 61, no. 2, pp. 103–113, 2018, doi: 10.1016/j.pcad.2018.06.004.

[27] C. M. Hernández-Ruiz, S. A. Villagrán Martínez, J. E. Ortiz Guzmán, and P. A. Gaona Garcia, "Model based on support vector machine for the estimation of the heart rate variability," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11140 LNCS, pp. 186–194, 2018, doi: 10.1007/978-3-030-01421-6_19.



Edvard Frederick Bareño Castellanos

Is a student in his final year of systems engineering in the faculty of engineering at Universidad Distrital Francisco José de Caldas (UDFJC) in Bogotá Colombia, graduated from the district school Antonio Nariño in 2011 in Bogotá Colombia. Since 2016 I am part of the GIIRA research group at Universidad Distrital Francisco José de Caldas (UDFJC). The areas of interest in research are artificial intelligence, artificial vision, neural networks, applied mathematics, programming and data analysis. Currently he works at Universidad Distrital Francisco Jose de Caldas in systems advisory office performing the functions as a junior developer in multiple systems of the university.

Paulo Alonso Gaona García

He earned a Ph.D. in Information and Knowledge Engineering from University of Alcalá in 2014 and working as a collaborator researcher in the Information Engineering Research Unit at University of Alcalá since 2012. Is full professor at Engineering Faculty of Universidad Distrital Francisco José de Caldas, Bogotá – Colombia since 2008. He is director of Multimedia Research Group and active member of GIIRA research group since 2008. He has a Master in Information Science and communication from Universidad Distrital Francisco José de Caldas (Bogotá - Colombia - 2006). He is Systems Engineer at Universidad Distrital Francisco José de Caldas (2003). His research interest includes web science, semantic web, network and communications, e-learning, information visualization and visual analytics.

Johan Enrique Ortiz Guzmán

I am Johan Enrique Ortiz Guzmán, professional in Sports Sciences, graduated from the University of Applied and Environmental Sciences U.D.C.A, in Bogotá Colombia, in 2009 and Master in Physiology, graduated from the University of Valencia, Spain, in 2011. I have been a university professor since 2012, working in faculties of medicine and health sciences such as the Universidad del Rosario, the UDCA university and the Fundación Universitaria del Area Andina, teaching classes in the area of physiology and sports control and integrating the area of basic sciences from the physiology unit. In terms of research, the focus of my work has been oriented towards cardiovascular adaptations to exercise, especially from the area of modulation exerted by the autonomic nervous system on cardiovascular function and this measured through the analysis of the heart rate variability.

Carlos Enrique Montenegro Marin

PhD in Computer Science from Oviedo University (2012). Master in Web Site Management and Engineering at the International University of the Rioja - UNIR (2013). Master in Information Science at the Universidad Distrital Francisco José de Caldas (2006). Systems Engineer at the District University (2003). He was a Dean of Engineering Faculty (December 2018 - 2019) and full professor, attached to Engineering Faculty of District University "Francisco José de Caldas" since 2006. He was Coordinator of committee of accreditation in the bachelor's in systems engineering (2007 to 2010), He was Coordinator of the bachelor's in systems engineering (2012 to 2014).

Eye-Tracking Signals Based Affective Classification Employing Deep Gradient Convolutional Neural Networks

Yuanfeng Li[1], Jiangang Deng[1], Qun Wu[2], Ying Wang[3]*

[1] Jiyang College of Zhejiang A&F University, Hangzhou (PR China)
[2] Institute of Universal Design, Zhejiang Sci-Tech University, Hangzhou (PR China)
[3] Department of Industrial Design at College of Art and Design, Zhejiang Sci-Tech University, Hangzhou (PR China)

Received 9 December 2020 | Accepted 9 April 2021 | Published 17 June 2021

Abstract

Utilizing biomedical signals as a basis to calculate the human affective states is an essential issue of affective computing (AC). With the in-depth research on affective signals, the combination of multi-model cognition and physiological indicators, the establishment of a dynamic and complete database, and the addition of high-tech innovative products become recent trends in AC. This research aims to develop a deep gradient convolutional neural network (DGCNN) for classifying affection by using an eye-tracking signals. General signal process tools and pre-processing methods were applied firstly, such as Kalman filter, windowing with hamming, short-time Fourier transform (SIFT), and fast Fourier transform (FTT). Secondly, the eye-moving and tracking signals were converted into images. A convolutional neural networks-based training structure was subsequently applied; the experimental dataset was acquired by an eye-tracking device by assigning four affective stimuli (nervous, calm, happy, and sad) of 16 participants. Finally, the performance of DGCNN was compared with a decision tree (DT), Bayesian Gaussian model (BGM), and k-nearest neighbor (KNN) by using indices of true positive rate (TPR) and false negative rate (FPR). Customizing mini-batch, loss, learning rate, and gradients definition for the training structure of the deep neural network was also deployed finally. The predictive classification matrix showed the effectiveness of the proposed method for eye moving and tracking signals, which performs more than 87.2% in accuracy. This research provided a feasible way to find more natural human-computer interaction through eye moving and tracking signals and has potential application on the affective production design process.

Keywords

Affective Computing, Deep Gradient Convolutional Neural Networks, Eye Moving and Tracking Signal, Short Time Fourier Transform.

DOI: 10.9781/ijimai.2021.06.002

I. Introduction

HUMAN affective comes from the physiological response of the machine to the periphery. While specific physiological activities generate different affective experiences, such as fear, may cause the accelerated heart rate, irregular breathing rhythm, abnormal skin electrical response, and the formation of a corresponding [1]-[3]. The brain mainly drives the dominant affective; under laboratory conditions, evoking an individual's real affective experience and acquiring physiological signals at the same time has been regarded as a popular affective measurement method currently. The affective generation method may be divided into material stimulation and situation induction [4]. Previous studies have shown that, under affective stimulation, obtaining human stress data through physiological signals and then analyzing it is a relatively stable and reliable method [5], including skin temperature, electromyography (EMG), electrocardiogram (ECG), blood volume, etc. [6]. To data, affective recognition based on multiple physiological signals is more complicated; firstly, the processing of physiological signals is more complicated, and the denoising algorithm needs to be developed stably continuously [7]. Weak noise may cause significant overall changes finally. Secondly, various physiological signals have unique denoising methods; common noises include motion artifacts, power frequency noise, and baseline drift caused by sensor movement. Besides, the qualitative problem of affective is currently more complicated, and it is mainly divided into discrete affective definitions and dimensional definitions. Either definition requires a self-assessment or manual labeling process, which brings a lot of subjectivity to the overall framework of affective recognition in this research. Now psychology does not conclude that the method is good, but from an experimental point of view, discrete affective are better recognized when the types of affective signal are few and clear. Finally, the physiological signals come from personal differences. In some scenes, there may be people who reflect strongly that some people have no fluctuations. The physiological signal is also that a person may have blood pressure

* Corresponding author.

E-mail address: wangying1980@163.com

and heartbeat that is already fast. There are two ways to solve this state. Among them, detecting the primary form in a calm state and regularizing the features in the affective state is acceptable [8].

This paper studies the recognition of human affective by eye movement signals, uses efficient preprocessing methods and feature calculations, and uses deep neural networks to complete the precise classification of affective. Eye movement signal extraction methods include human eye pupil positioning and eye movement feature extraction; among them, techniques such as pupil center positioning based on gray-scale information are performed. The eye movement signal is a signal of potential changes around the eye generated by eye movement. The signal has the advantages of high amplitude, easy waveform recognition, and simple processing. At present, eye-tracking signals for human-computer interaction research are a hot issue and a relatively new direction [9]-[11]. Some scholars studied people's mental activities by examining their eye movements and explored the relationship between eye movements and human mental activities by analyzing the recorded eye movement data. The eye tracker's advent provides a new way for psychologists to use eye moving and tracking techniques to explore the visual information processing mechanism of humans under various conditions and observe their wonderful or interesting relationship directly or indirectly with mental activities [12] [13]. The eye movement technology has experienced observation methods, post-image methods, mechanical recording methods, optical recording methods, image recording methods, and other methods.

Eye moving and tracking technology extracts data features such as fixation point, fixation time and frequency, saccade distance, pupil size, etc. From the recording of eye movement trajectory, to study an individual's internal cognitive process, Juhola et al. used eye movement signal analysis for otoneurologic patients [14][15]; Kasneci, E., et al. employed eye-tracking and aggregated physiological signals for perception prediction [16]. Eye-tracking also may be applied for computer-aided diagnosis [17], equipment operator [18], and video learning [19][20]. The eye-tracking signals were also successfully applied to detect autism patients [21]-[23]. There are three basic ways of eye movement: fixation, saccades, and pursuit movement. Eye movement reflects the selection mode of visual information, which great impacts discovering the psychological mechanism of cognitive processing. From the research report, the commonly used data or parameters of psychological research using eye trackers mainly include gaze point trajectory map, eye movement time, average velocity, amplitude time and distance, pupil size (area or diameter, unit pixel) and blink. The spatiotemporal feature of eye movement is the physiological and behavioral performance in visual information extraction. It has a direct or indirect relationship with human psychological activities, so many psychologists are devoted to eye movement research. When the eyeball moves, a weak magnetic field is formed between the retina and the cornea. According to the study, the potential change between the cornea and the retina is highly correlated with the eyeball's rotation angle. It is linearly correlated between 0-30, and 30-60 is the relationship between the sine and cosine line. Electrodes may not be directly assigned to the cornea for measurement while skin electrical tests are performed on the eye's outer skin. But this is different from surface EMG (sEMG), for the test range and signal are weak and stable information cannot be obtained. Therefore, the currently effective method is to set a mirror on the cornea or iris for optical measurement. Among them, the contact eyepiece is reflected light. The other method is the detection coil method, which determines the direction of the eye movement by the change of the magnetic field around the eye. At present, electro-oculography (EOG) signal method is also popular [24].

A deep neural network was applied for classifying eye movement signals recently. The simplest artificial neural network is a binary linear classifier; a neuron's structure can be divided into dendrites, synapses, cell bodies, and axons. A single neuron can be regarded as a machine with only two states. The transformation of a neuron depends on the number of input signals received from other neurons and the synapses' strength. When the semaphores' sum exceeds a certain threshold, the neuron body will be excited and generate electrical pulses [25][26]. Electrical pulses are transmitted along the axon and through the synapse to other neurons which is defined as a synapse, bias as a threshold, and activation function as neuron body. The deep network of unsupervised learning is aimed at pattern analysis or synthesis tasks, capturing high-order correlations of observed or visible data without target label information; and the supervised learning deep network directly provides the discriminative ability for pattern classification. Describes the posterior distribution under visible data, also called discriminative deep network (DDN). Besides, there is another type called hybrid deep network (HDN) based on a discriminative model. The unsupervised deep network mainly includes deep Boltzmann machine, sum-product networks (SPN), recurrent neural network (RNN), etc. The training process of deep neural networks is critical in the process of signal classification. How to make the network perform well on the training set and make the network perform the same on the test set if the training set performs well is critical. How to tune on the training set is also a vital topic, including how to choose the appropriate loss function, mini-batch, choose a new activation function, adaptive learning rate, and momentum. The current practice is to increase the training set, stop early, regularize, dropout, and improve network structure [27]. The deep convolutional neural network (DCNN) has developed rapidly and was initially used for image recognition classification [28]. Using the convolution layer and the pooling layer, the ability to accurately predict the image is accelerated. RNN and its derived algorithms can be used for speech recognition, natural language processing, speech synthesis, etc. The deep neural network (DNN) can model changes in time series. The research of DNN for classification is relatively mature, such as LeNet, AlexNet, ZFNet, VGGNet GoogLeNet, Inception-v1, etc. [29]-[31]. Applied eye movement signals for affective classification is feasible due to the deep learning technologies applied in other datasets, such as images [32] and wide applications in industrial engineering [33] [34].

This paper developed a novel deep neural network based on gradients calculation by converting eye moving and tracking signals to images; feature extraction and evaluation indices were also defined for the comparing analysis and finally performed the proposed model for the dataset. The organization of the remaining sections is as follows. Section II introduces modeling for the dataset, including preprocessing algorithms. Section III addresses the results and comparing analysis. Section IV involves the concluding remarks of the studies and future works.

II. Modeling

A. Preprocessing

The specific requirements for signal preprocessing are due to the features of the vibration signal itself. The function of signal preprocessing is to make a certain extent by using important factors for influencing subsequent signal analysis. The fast Fourier transform (FFT) is a general term for efficient and fast calculation methods through calculating discrete Fourier transform (DFT). DFT can discretize the finite-length sequence in the frequency domain, but its calculation is too large to handle the problem in real-time; the basic idea of FFT is to sequentially decompose the original N-point sequence into a series of short sequences [35][36]. Make full use of the symmetric and periodic nature of the exponential factors in the DFT calculation formula, and then find the corresponding DFT of these short sequences and make appropriate combinations to achieve the purpose of eliminating duplicate calculations, reducing multiplication operations, and simplifying the structure. Fast algorithms such as high basis and split basis have been developed, such as the Winograd Fourier transform algorithm (WFTA) [37] and prime factors based on number theory and polynomial theory Fourier transform algorithm. Their common feature is that when N is a prime number, the DFT calculation can be converted into a circular convolution, reducing the number of multiplications and increasing the speed [38]. The obvious advantage of a small calculation amount makes FFT widely used in signal processing technology, and real-time processing of signals can be realized in combination with high-speed hardware.

Another preprocessing method is signal filtering (i.e., wave filtering), an operation to filter out specific frequency bands in the signal and is an important measure to suppress and prevent interference. Filtering is divided into classic filtering and modern filtering. Classical filtering is an engineering concept based on Fourier analysis and transformation. According to higher mathematics theory, any signal that satisfies certain conditions can be regarded as a superposition of infinite sine waves. In other words, the signal is a linear superposition of sine waves of different frequencies. The sine waves of different frequencies. The sine waves of different frequencies that make up the signal are called the signal's frequency component or the harmonic component. The Kalman filter used in this study is a recursive algorithm (i.e., real-time algorithm). Specifically, for discrete-time filtering, as long as the dimension of X is appropriately increased, the filter value table at time t can be some linear combination of the filter value at the previous time and the observation value $Y(t)$ at the current time. For continuous-time filtering, the linear stochastic differential equations that should be satisfied with $Y(t)$ may be given. In the case where the observation results and output filter values need to be continuously increased, such an algorithm speeds up the processing of data and reduces the amount of data storage.

First, the following filter equations need to be used for determining parameters here. The Kalman filter is suitable for linear systems, and the state equation and observation equation of the system are,

$$x(k) = Ax(k-1) + Bu(k) + w(k) \tag{1}$$

$$z(k) = Hx(k) + y(k) \tag{2}$$

Where, $x(k)$ is the state of the system at time k; $u(k)$ is the control quantity of $x(k)$; $w(k)$ is to process the noise conforming to the Gaussian distribution; and the is that, $z(k)$ is the observation value of the system at time k while $y(k)$ measures the noise conforming to the Gaussian distribution. The covariance is A, B and H present the system parameters, matrix for multiple input and multiple output, and several constants for single input and single output separately. The algorithm 1 describes the filter in detail [39].

```
Algorithm 1: Kalman Filter
Require: data: input signal; Q: covariance of w(k); R: covariance of z(k); x0; p0.
Output: P
L ← length(data) # calculate the length of the input data
K ← zero(L,1) # set the initial zeros to the parameters
X ← zero(L,1)
P ← zero(L,1)
X (1) ← x0; # set the start point
P (1) ← p0
FOR i in (2 to L)
    K(i) ← P(i-1) / (P(i-1) + R)
    X(i) ← X(i-1) + K(i) * (data(i) - X(i-1))
    P(i) ← P(i-1) - K(i) * P(i-1) + Q
ENDFOR
```

Another preprocessing algorithm is the short-time Fourier transform (STFT), a variant of the Fourier transforms, also known as windowed Fourier transform or time-dependent Fourier transform, used to determine the sinusoidal frequency and phase of a local portion of a signal that changes with time. The process of calculating the STFT is to divide the long-term signal into several shorter equal-length signals and then calculate the Fourier transform of each shorter segment separately. It **usually describes** changes in the frequency and time domains and as one of the **essential** tools in time-frequency analysis [40]. Continuously, we have that,

$$X(f) = \int_{-\infty}^{\infty} \frac{x(t)}{e^{2\pi ift}}\, dt \tag{3}$$

$$X(t,f) = \int_{-\infty}^{\infty} \frac{\varpi\,(t-\tau)x(\tau)}{e^{2\,\pi\, if\,\tau}}\, d\tau \tag{4}$$

Where, f is frequency of the signal; t is time. The STFT is to truncate the original Fourier transform into multiple segments in the time domain and perform the Fourier transform separately. Each segment is recorded as time t_i, and the frequency domain characteristics are obtained by corresponding FFT, and the time t_i can be roughly estimated. The frequency domain characteristics of time (that is, the corresponding relationship between the time domain and the frequency domain is simultaneously guided). The tool used for signal truncation is called a window function (the width is equivalent to the length of time). The smaller the window, the more obvious the time-domain characteristics, but currently, for the number of points is too small, the FFT reduces the accuracy and the frequency-domain characteristics are not obvious. In order to ensure the improvement of the time domain characteristics on the basis of the frequency domain characteristics, it is often selected to overlap a part of the front and rear window functions, so that the time of the two windows is determined is closer to improve the time domain analysis ability. However, it is not better to have more overlaps. Too many overlapping points will greatly increase the amount of calculation which is resulting in low efficiency. Therefore, the number of overlapping points in the front and rear windows also needs to be determined. Several windowed models are introduced as below,

- Hamming window:

$$w(n) = 0.54 - 0.46\cos\left(\frac{n}{N}\right) \tag{5}$$

- Hanning window:

$$w(n) = \frac{\left(1 - \cos\left(\frac{n}{N}\right)\right)}{2} \tag{6}$$

- Rectangular window

$$w(n) = 1.0 \tag{7}$$

- Triangle window

$$w(n) = TRI\left(\frac{2n}{N}\right) \tag{8}$$

- Blackman, third order raised cosine window

$$w(n) = 0.42 + 0.5\cos\left(\frac{n}{N}\right) + 0.08\cos\left(\frac{2n}{N}\right) \tag{9}$$

- Blackman-Harris window

$$w(n) = 0.359 - 0.489\cos\left(\frac{n}{N}\right) + 0.141\cos\left(\frac{2n}{N}\right) - 0.012\cos\left(\frac{3n}{N}\right) \tag{10}$$

In this research, the window function selects the hamming window, and the maximum number of DFT points is not greater than 256; user input (pass value) is signal, window, overlap, N, fs, etc. According to the size of the window, split the signal and multiply it with the window function; perform N-point FFT on each signal segment and find the energy spectral density; Algorithm 2 described the windowing process on eye movement signals, and splitting the signals for further analysis was described in Algorithm 3 [41].

Algorithm 2 preprocessing the eye movement signal by windowing

Required: signal.vector: original signal sequence; N: length of signal; w: window length; noverlap: the number of overlapping windows; nfft: FFT/DFT points; fs sampling frequency;

Output: signal

```
FOR i in (0 to w) #hamming window
      hamming[i] = 0.54 - 0.46 * cos(2*M_PI*i/(w-1))
      windowPV ⟵ windowPV+ pow(hammingW[i], 2)
ENDFOR
#calculate the number of rows and columns of the short-time Fourier transform signal array, the number of rows is the number of time points, the number of columns is w
row ⟵ (N - noverlap)/ (w - noverlap)
column ⟵ w
half_Nfft ⟵ nfft/2+1
FOR I in (0 to row)
      timeV[i] ⟵ ((float)i) /((float)(fs*(w/2+1+(w-noverlap) *i)))
ENFOR
# separating the signals by row and column using windowing
FOR i in (0 to row)
     FOR j in (0 to column)
            signalXY[i][j] ⟵ signal.vector[i*(w - noverlap) + j]
            signalXY[i][j] ⟵  signalXY[i][j] *hammingW[j]
     ENDFOR
ENDFOR
```

B. Feature Extraction Methods

The feature extraction (FE) methods include time domain, frequency domain and model-based method. For time-domain based features, there are waveform, pulse, kurtosis, margin, peak and zero crossing rate (ZCR), etc. the ZCR is calculated by,

$$ZCR_n = \frac{1}{2}\sum_{i=0}^{N-1}|\mathrm{sgn}[x_n(i)] - \mathrm{sgn}[x_n(i-1)]| \tag{11}$$

Where, N is the length of the frame, n is the number of frames; $\mathrm{sgn}(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0. \end{cases}$

Using time as an independent variable to describe changes in physical quantities is the most basic and intuitive expression of signals. In the time domain, the signal is filtered, amplified, statistical feature calculation, correlation analysis and other processing, collectively called the time domain analysis of the signal. Different eye movement waveforms have obvious differences. It is also feasible to directly analyze the waveform characteristics, such as amplitude and wavelength. In addition, the digital signal can also be fitted by the model method. As commonly used linear predictive coding (LPC). For eye movement signal, frequency domain-based feature methods mainly focus on power spectrum density (PSD) estimation. Here, we adopted parametric method, supposed that., eye movement signal is $x(n)$, the autocorrelation function of $x(n)$ is r(k), so, the PSD is,

$$p(w) = \sum_{k=-\infty}^{+\infty} r(k)e^{-jwk} \tag{12}$$

Algorithm 3: FFT and STFT preprocessing for eye moving and tracking signals

Required: nfft: FFT/DFT points; fs: sampling frequency

Output: S; P;

```
freqStep ← fs/nfft; freq[0] ← 0.0
  FOR i in (1 to nfft)
       freq[i] ← freq_Step + freq[i-1]
  ENDFOR
  FOR i in (0 to row)
    FOR j in (0 to half_Nfft)
       Sxx[i][j] = (fft_Real[j]*fft_Real[j] + fft_Img[j]*fft_Img[j])/ windowPV
    ENDFOR
  ENDFOR
  Pxx[0][0] ⟵ Sxx[0][0]/fs
  log_Pxx[0][0] ⟵ 10*log10f(fabsf(Pxx[0][0]))
  pxx_Max ⟵ log_Pxx[0][0]
  pxx_Min ⟵ log_Pxx[0][0]
 FOR i in (1 to row)
      Pxx[i][0] ⟵ Sxx[i][0]/fsFloat
      log_Pxx[i][0] ⟵ 10*log10f(fabsf(Pxx[i][0]))
      IF log_Pxx[i][0] > pxx_Max THEN
          Pxx_Max ← log_Pxx[i][0]
      ENDIF
      IF log_Pxx[i][0] < pxx_Min THEN
          pxx_Min ← log_Pxx[i][0]
      ENDIF
    IF nfft mod 2 = = 0 THEN
       FOR i in (0 to row)
           FOR j in (1 to half_Nfft)
               Pxx[i][j] ← Sxx[i][j] *2.0/fs
               log_Pxx[i][j] ← 10*log10(abs (Pxx[i][j])) # abs(.):absolute
               IF log_Pxx[i][j] > pxx_Max THEN
                    pxx_Max ← log_Pxx[i][j]
               ENDIF
               IF log_Pxx[i][j] < pxx_Min THEN
                    pxx_Min ← log_Pxx[i][j]
               ENDIF
           ENDFOR
       ENDFOR
     ELSE
        FOR i in (0 to row)
              Pxx[i][half_Nfft-1] ← Sxx[i][half_Nfft-1]/((float)fs)
              log_Pxx[i][half_Nfft-1] ← 10*log10(abs (Pxx[i][half_Nfft-1]))
              IF log_Pxx[i][half_Nfft-1] > pxx_Max THEN
                   pxx_Max ← log_Pxx[i][half_Nfft-1]
              ENDIF
        ENDFOR
        FOR i in (0 to row)
             FOR j in (1 to halfNfft-1)
                  Pxx[i][j] ← Sxx[i][j] *2.0/((float)fs)
                  log_Pxx[i][j] ← 10*log10f(fabsf(Pxx[i][j]))
                  IF log_Pxx[i][j] > pxx_Max THEN
                       pxx_Max ← log_Pxx[i][j]
                 ENDIF
                 IF log_Pxx[i][j] < pxx_Min
                      pxx_Min ← log_Pxx[i][j]
                 ENDIF
             ENDFOR
        ENDFOR
   ENDIF
 ENDFOR
```

Where, $r(k) = EP[x(n)x * (n+k)]$,* is conjugate; EP is expectation.

If the length of eye movement signal is smaller, the PSD may be adjusted by,

$$\hat{P}(w) = \sum_{k=-(N-1)}^{N} r(k)e^{-jwk} \tag{13}$$

Where, $\hat{r}(k) = \frac{1}{N-k}\sum_{n=k+1}^{N} x(n)x * (n-k)\,, 0 \le k \le N-1$.

Parameter spectral density estimation first uses a certain function to fit the signal, then you can determine each parameter of the function model, and finally get the spectral characteristics of the parameter. When the signal fitted by the established function model is like the actual signal, the efficiency of the parameter spectral density will be relatively high. The commonly used models of modern spectral density are automatic moving average (ARMA), auto regressive (AR), moving average (MA), etc. For eye movement models, AR models are often used. The main advantage is that under high SNR conditions, a relatively high resolution can be obtained. It is suitable for short data processing. AR is a linear regression model that linearly combines random variables at several times in the early period to describe random variables at a certain period in the later period. It is essentially a linear prediction, expressed as follows:

$$x(n) = -\sum_{i=1}^{c} a_c(i)x(n-i) + \varepsilon(n) \tag{14}$$

Where, $\varepsilon(n)$ is a white noise sequence with variance σ^2 and mean is zero, and c represents the order of the AR model. Therefore, the eye movement signal sequence $x(n)$ may be regarded as the output of the white noise sequence $\varepsilon(n)$ through the AR model. When building an AR model, the first question is to determine the appropriate order. The order of the model is implemented in the recursion process. When using the LevinsonDurbin recursion method (LRM), each set of parameters from low order to high order can be given. When the minimum prediction error power of the model no longer changes, the correct order is acquired.

In this research, frequency domain indicators are also applied including center of gravity frequency (CGF), mean square frequency (MSF), root mean square frequency (RMSF), frequency variance (FV), frequency standard deviation (FSD), short-term power spectral density (STPSD), spectral entropy (SE), fundamental frequency (FF), and formant (F). We combined all signal features and used transform processing technologies to prepare the inputs for deep neural networks. The candidate transforms include FFT, non-parametric power spectrum (i.e., periodic graph method, Welch method), parametric power spectrum estimation method, STFT, wavelet transform, Hilbert transform, MFCC, Wigner distribution (WDF), Radon transform, and Gabor transform.

Algorithm 4 deep gradients convolutional neural network for eye movement signals classification

```
Required: numObservations: number of observations; miniBatchSize; maxEpochs; numIterationsPerEpoch
Output: loss; learnRate; epoch; Elapsed time
Iteration ← 0
numIterationsPerEpoch ← floor(numObservations./miniBatchSize);
start ← tic
FOR epoch in (1 to maxEpochs)
  idx ← randperm(numObservations)
  XTrain ← XTrain(idx)
  YTrain ← YTrain(idx)
  FOR i in (1: numIterationsPerEpoch)
    iteration ← iteration + 1
    #Read mini-batch of data and apply the transformSequences
    idx ← (i-1)*miniBatchSize+1:i*miniBatchSize;
    [X,Y,numTimeSteps] ←
    transformSequences(XTrain(idx),YTrain(idx));
    dlX ← dlarray(X)
    # Evaluate the model gradients and loss using dlfeval.
    [gradients, loss] ←
          dlfeval(@modelGradients,
          dlX,Y,parameters,hyperparameters,numTimeSteps)
    #Clip the gradients.
    gradients ← dlupdate(@(g)
    thresholdL2Norm(g,gradientThreshold),gradients);
    # Adam optimizer for updating the network
    [parameters,trailingAvg,trailingAvgSq] ←
    adamupdate(parameters,gradients, ...
      trailingAvg, trailingAvgSq, iteration, learnRate)
    IF plots == "training-progress"
      D ← duration (0,0, toc(start),'Format','hh:mm:ss')
      loss ← mean (loss/ numTimeSteps)
      loss ← double(gather(extractdata(loss)))
      loss ← mean(loss)
      addpoints(lineLossTrain,iteration, mean(loss));
      title ("Epoch: " + epoch + ", Elapsed: " + string(D))
      drawnow # plot the results
    ENDIF
  ENDFOR
  IF mod(epoch,learnRateDropPeriod) == 0
    learnRate = learnRate*learnRateDropFactor;
  ENDIF
ENDFOR
```

C. Deep Gradients Convolutional Neural Network-based Classification Model

The gradient is a concept related to the directional derivative. The direction of the gradient is given by the angle of the gradient vector relative to the x-axis. The original gradient descent algorithm is obtained from the directional derivative, and then the momentum gradient descent algorithm is described. Due to the importance of hyperparameter learning rate to gradient descent, the gradient algorithm has multiple adaptive gradient descent algorithms; the forms of gradient descent include batch gradient descent (BDG), stochastic gradient descent (SGD), and mini-batch gradient descent (MGD); the evolution of gradient descent is mainly several adaptive gradient descent algorithms such as AdaGrad, RMSprop, AdaDelta, and Adam. Gradient descent is a commonly used optimization method for machine learning. The difficulty of gradient descent is mainly reflected in the setting of the learning rate.

The minimum point, the first difficulty in saddle point gradient descent optimization, is the problem of setting the learning rate mentioned above. When the learning rate is too low, the convergence speed is slow, and when the learning rate is too large, it will cause training shocks and may diverge. In contrast, non-convex error functions generally appear in neural networks. When optimizing such functions, another difficulty is that the gradient descent process may fall into a local minimum. Studies have also pointed out that this

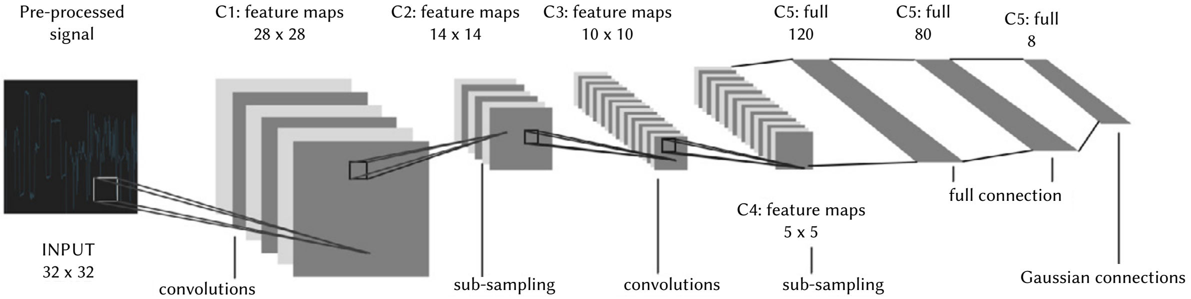


Fig. 1. deep gradients convolutional neural network for eye movement signals classification.

difficulty does not come from the local minimum, but more from the saddle point, those that are increasing in one dimension and decreasing in another dimension. Points with the same error usually surround these saddle points. Because the gradient in any size is approximately 0, it is difficult for SGD to escape from these saddle points.

The neural network's main task is to find the optimal parameters (weights and biases) during learning. This optimal parameter is also the parameter when the loss function is minimal. However, in general, the loss function is more complicated, and there are many parameters, so it is impossible to determine where to obtain the minimum value. So, the gradient method is the method to find the minimum value (or as small as possible) through the gradient. It is noted that the gradient indicates the direction in which the function value at each point decreases the most, so the path of the gradient does not necessarily point to the minimum. But along its direction can minimize the value of the function. Therefore, when looking for the minimum value (or as small as possible) of the function, use the information of the gradient as a clue to determine the direction of progress. Currently, the gradient method comes in handy. In the gradient method, the value of the function advances a certain distance from the current position along the gradient direction, then recalculates the gradient in the new direction, and then advances along the new gradient direction, and so on. Like this, the process of gradually decreasing the value of the function by continuously advancing in the direction of the gradient is the gradient method (gradient method). In neural networks (deep learning), the gradient method mainly refers to the gradient descent method [39]. We designed a deep gradients convolutional neural network-based classification model, which is illustrated in Fig.1.

Algorithm 4 introduced the process of the network proposed.

III. Results and Discussion

A. Data Acquisition and Preprocessing

Normally, the amplitude of the eye movement signal is in the range of 0.4-10mv, and the frequency is between 0-38Hz; the main frequency is 0-10Hz. While in this experiment, the Sampling frequency used for eye tracking is 600Hz. Eye image stream frequency is approximately 10 Hz. Accuracy is 0.3° at optimal conditions (down to 0.16°); 8-bit timestamped data (256 event codes) event-driven detection with a timestamp accuracy of 50 μs; Operating distance is of 55 to 75 cm from the eye tracker reference point; freedom of head movement is 34 cm width x 26 cm height at 65 cm, that means at least one eye tracked. Two cameras capture stereo images of both eyes for accurate measurement of eye gaze and position in space. The experiment uses a "non-invasive" technology based on video oculographic (VOG),. The specific model is Tobii's X6, 0/120 series of bare machines; the user's range of activities is limited to a square of 2 meters, and authorized test personnel are required to place their chin on a fixed bracket to obtain more accurate data. The system compensates the head movement, and the specified head movement range is 44×22×30cm (length, width and height). This experiment collected 16 group signals (8 males, 8 females, 18-22 years old). Fig. 2 shows the original tracking patterns, in which the black dots are the location, lines between dots are the traces. Fig. 3 shows the signals processed by using SIFT. FFT applied for eye-tracking signals is acquired from the experimental platform. (shown in Fig. 4)

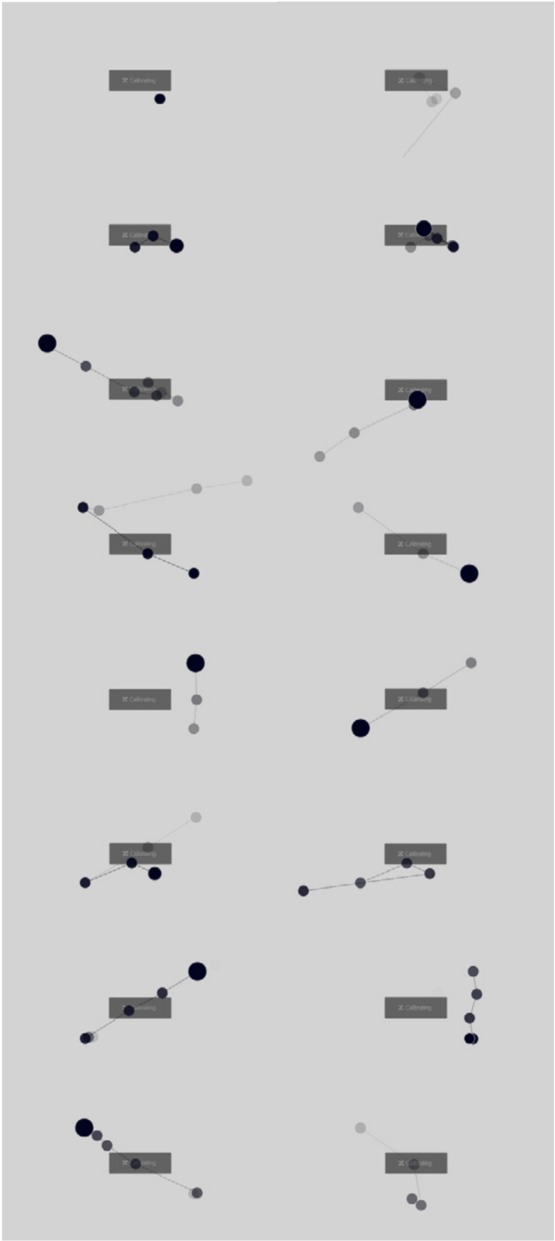


Fig. 2. Eye movement tracking scene experiment.

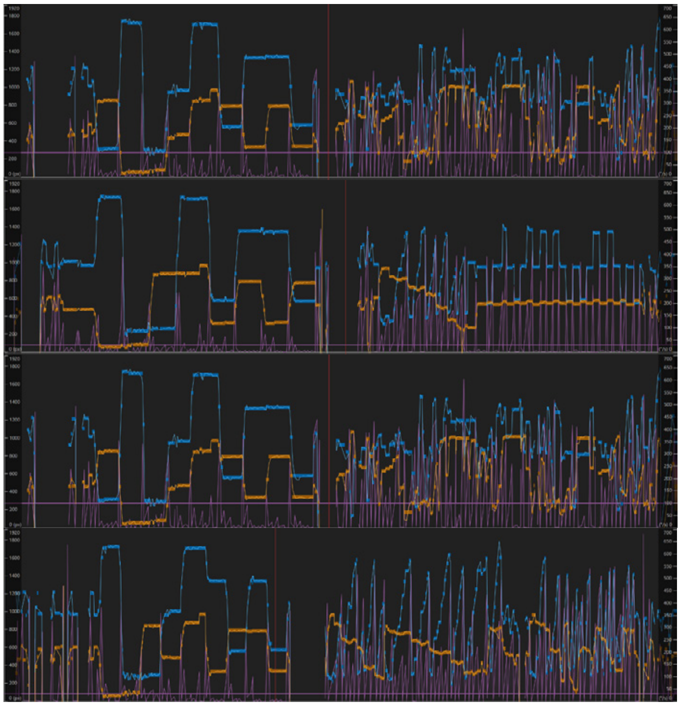


Fig. 3. Eye movement preprocessed signals acquired by eye tracking lab system Tobii pro.

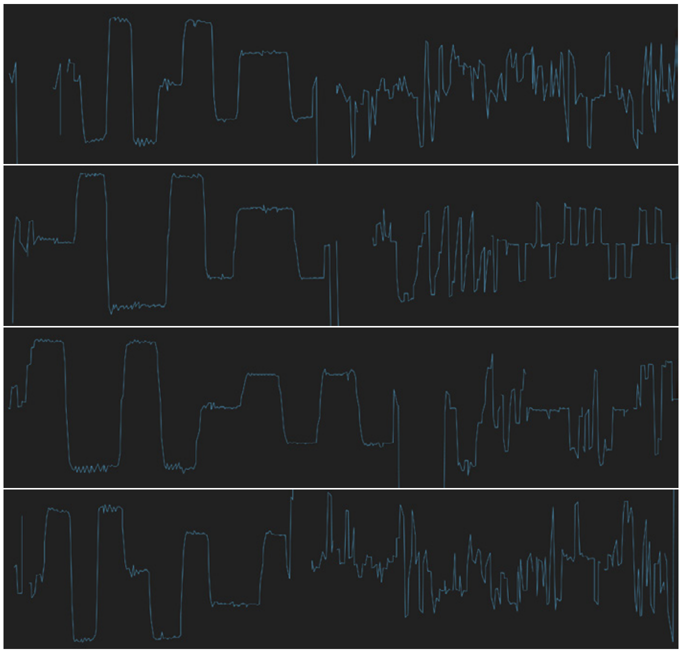


Fig. 4. FFT applied for Eye movement signals.

B. Classification Results and Discussion

Frequency, sample frequency, eye gaze, eye position, moving speed, peak value of signals, time-synchronized average (TSA), amplify, root mean square, Impulse Factor, Signal-to-Noise Ratio (SNR), Total Harmonic Distortion (THD). The input features dataset is listed in Appendix A. The original training dataset for eye movement signals with features and manual labels are illustrated in Fig. 5. The features were labeled by a group of person in advance using investigation system. We have the 600 rows with 13 columns which including 12 features and 1 label for 4 affective of nervous, calm, happy, and sad. We also developed an evaluation index table for the performance of the classification model. TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Then we have that, Accuracy (A) is for the entire model, the number of prediction pairs is higher than the total number of samples. Accuracy is TP+TN/(TP+TN+FP+FN). Precision (P) is the prediction is correct for a certain target, which is compared to the total number of that type predicted, P is TP/(TP+FP). Recall (R) or recall rate is for a certain category that how much of the category has been predicted in total. R is TP/(TP+FN). F1-score is for the case of the precision rate and the recall rate are always opposites; F1-score is an indicator that can balance the two to choose an optimal value. F1 is 2(PR)/(P+R). The proposed DGCNN classification results are shown in Fig. 6, which is the accuracy predictive matrix and AOC for DGCNN classification model and Fig. 7 shows the ROC curvature for the TPR and FNR by selecting affective of "calm"; the average accuracy of effective is more than 87.2%. Fig. 8 shows the predictive class matrix of decision tree (DT) model for eye-tracking signals; Fig. 9 shows the predictive class matrix of Gaussian naïve Bayes, and Fig. 10 shows the predictive class matrix of KNN model.

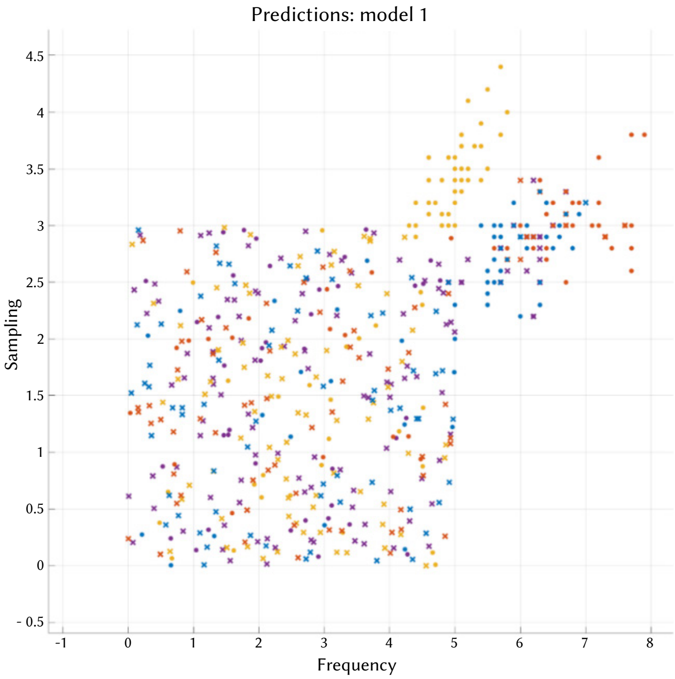


Fig. 5. The original training dataset for eye movement signals with features and manual labels.

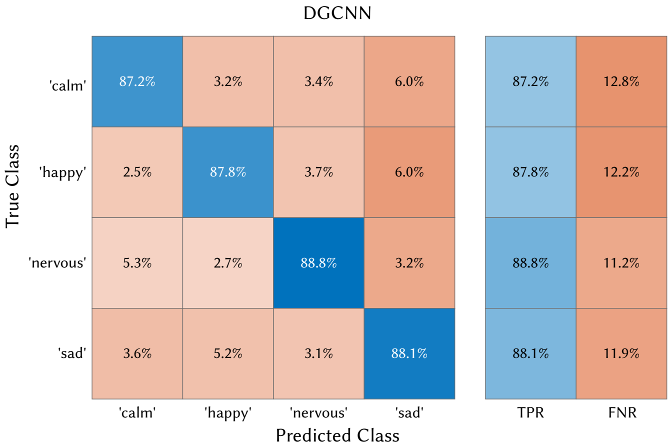


Fig. 6. Accuracy predictive matrix and AOC for DGCNN classification model.

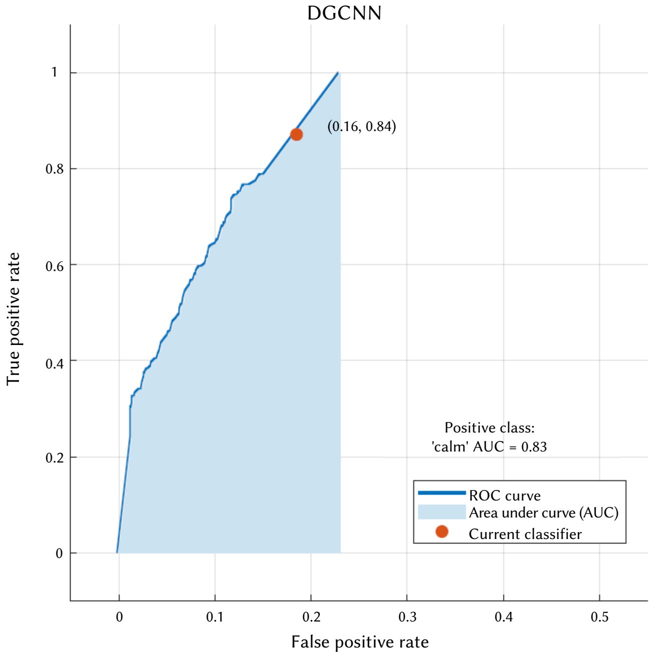


Fig. 7. ROC curvature for the TPR and FNR (affective =calm).

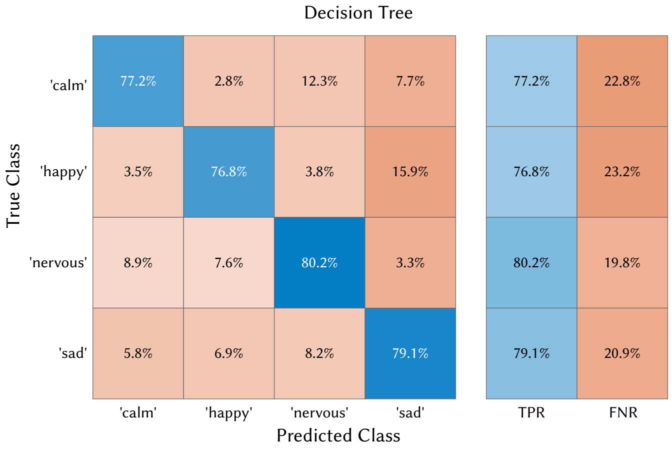


Fig. 8. Predictive class matrix of decision tree model for eye movement and tracking signals.

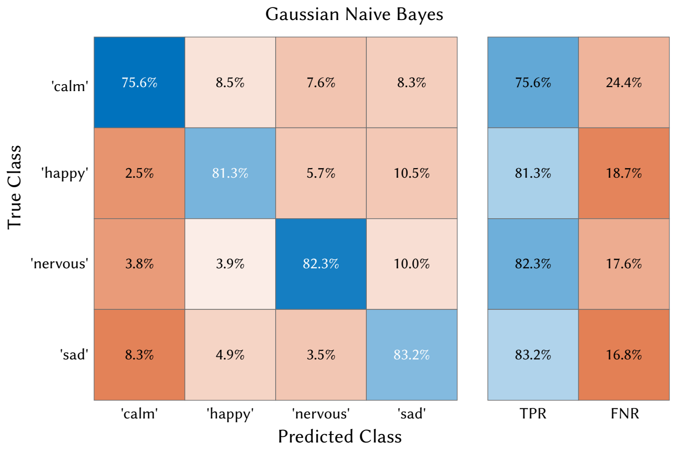


Fig. 9. Predictive class matrix of Gaussian naïve Bayes model for eye movement and tracking signals.

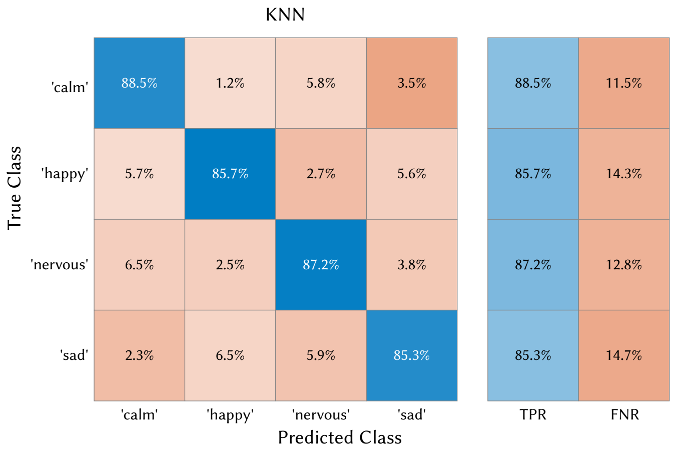


Fig. 10. predictive class matrix of KNN model for eye movement and tracking signals.

The matrix known that the proposed DGCNN performs a much higher predicted class in TPR and lower in FNR. Furthermore, more indices are developed for comparing the proposed model for classifying affective signals in which, A is accuracy; S1 is sensitivity; S2 is specific; PP is positive predictive; NP is negative predictive; PL is positive likely; NL is negative likely; F1 is f1-score which defined based on TPR and NPR. The comparing results are shown in Table I.

TABLE I. Comparative Analysis on the Different Classifiers and the Proposed Classification Model for Acoustic Signal Analysis with STFT Preprocessed Signals

Classifier/Evaluation Metrics	A	S1	S2	PP	NP	PL	NL	F1
ANN [43]	76.40%	81.23%	77.43%	78.12%	77.12%	7.21	0.13	0.91
MCSVM [44]	78.76%	82.12%	76.43%	76.32%	75.32%	7.22	0.12	0.82
LSTM [45]	81.42%	82.33%	76.78%	79.34%	77.32%	7.21	0.12	0.87
R-CNN [46]	82.43%	84.65%	78.76%	81.32%	78.12%	6.31	0.14	0.88
Fast R-CNN [47]	85.54%	84.43%	82.45%	82.57%	81.11%	6.54	0.12	0.87
VGG-16 [48]	85.43%	83.87%	84.67%	79.89%	81.81%	6.66	0.14	0.90
GoogLeNet [49]	87.11%	84.65%	83.55%	84.90%	81.78%	7.54	0.14	0.92
AlexNet [50]	85.76%	82.48%	85.12%	83.43%	82.32%	7.35	0.14	0.93
DGCNN (*)	88.10%	85.98%	82.69%	84.09%	84.87%	7.55	0.12	0.94

And the proposed DGCNN has still in high effectiveness in A, S1, NP, PL, and F1. Table I shows the comparing results for those indices, in which the DGCNN performs high effectiveness in most indices.

IV. Conclusion Remarks

This article mainly studies the acquisition and preprocessing of eye movement signals, endpoint detection, effective eye movement signal extraction, feature extraction, and compares classification algorithms such as SVM, Bayes, and GoogleNet to verify the effectiveness of this method; eye movement analysis method is an effective method for studying human cognitive processing. Its advantages of harmless, ecological and high-efficiency are difficult to replace by general research technology; eye-tracking technology has become the foundation of psychology, neuromarketing, neurocognition, user experience technical methods for visual behavior and human behavior in many fields such as research and market research, recently, deep CNN was also used in sentiment analysis and emotion detection in conversations [51]. The research limitation is that the eye-tracking signals need to be converted and lose some information in feature extraction. Furthermore, future research requires multimodal physiological signals; on the other hand, more human emotion classifications also need to be further evaluated.

Eye movement signals can better characterize the emotional state and perform efficient classification; in this paper, the eye-tracking signals can be transformed into the standard input of the convolutional neural network by utilizing the eye movement signals' transformation and preprocessing. The effectiveness of the method is verified by comparison analysis. In future works, affective computing will be widely applied to medical rehabilitation, and assisting autistic people for their emotional changes, applying affective computing in education to realize the collection and analysis of learning status and guiding the selection of content progress. The computer can perceive the user's preference for music and based on the understanding and judgment of the emotional response, provide users with more interesting music playback.

References

[1] Rahal, R.-M. and S. Fiedler, "Understanding cognitive and affective mechanisms in social psychology through eye-tracking". *Journal of Experimental Social Psychology,* vol. 85, pp. 103842, 2019, doi: 10.1016/j.jesp.2019.103842

[2] Kale, G. V., & Patil, V. H. "A study of vision based human motion recognition and analysis". *International Journal of Ambient Computing and Intelligence*, vol. 7, no.2, pp.75-92, 2016, doi: 10.4018/IJACI.2016070104

[3] Gargava, P., & Asawa, K. Brain, "Computer Interface for Micro-controller Driven Robot Based on Emotiv Sensors", *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 4, no.5, pp.39-43, 2017, doi: 10.9781/ijimai.2017.457

[4] Poria, S., Cambria, E., Bajpai, R., & Hussain, A., "A review of affective computing: From unimodal analysis to multimodal fusion", *Information Fusion,* vol. 37, pp. 98-125, 2017, doi: 10.1016/j.inffus.2017.02.003

[5] Lee, W., & Norman, M. D. "Affective Computing as Complex Systems Science", *Procedia Computer Science*, vol. 95, pp. 18-23, 2016, doi: 10.1016/j.procs.2016.09.288

[6] Belkacem, A.N., et al., "Erratum to Online classification algorithm for eye-movement-based communication systems using two temporal EEG sensors", *Control, Biomedical Signal Processing and Control*, vol. 19, pp. 137, 2015, doi: 10.1016/j.bspc.2015.01.006

[7] Dai, W., Han, D., Dai, Y., et al., "Emotion recognition and affective computing on vocal social media", *Information & Management*, vol. 52, no.7, pp.777-788, 2015, doi: 10.1016/j.im.2015.02.003

[8] Choi, M., Seo, M., Lee, J. S., et al., "Fuzzy support vector machine-based personalizing method to address the inter-subject variance problem of physiological signals in a driver monitoring system", *Artificial Intelligence in Medicine,* vol. 105, pp.101843, 2020, doi: 10.1016/j.artmed.2020.101843

[9] R.C.A. Bendall, S. Lambert, A. Galpin, et al., "A cognitive style dataset including functional near-infrared spectroscopy, eye-tracking, psychometric and behavioral measures", *Data in Brief*, vol. 26, 104544, 2019, doi: 10.1016/j.dib.2019.104544

[10] John Brand, Travis D. Masterson, et al, "Measuring attentional bias to food cues in young children using a visual search task: An eye-tracking study", *Appetite*, vol. 148, 104610, 2020, doi: 10.1016/j.appet.2020.104610

[11] Wen-ying Sylvia Chou, Neha Trivedi, et al., "How do social media users process cancer prevention messages on Facebook? An eye-tracking study", *Patient Education and Counseling*, vol. 103, no.6, pp.1161-1167, 2020, doi: 10.1016/j.pec.2020.01.013

[12] Gisele C. Gotardi, Sérgio T. Rodrigues, Fabio A. Barbieri, et al.," Wearing a head-mounted eye tracker may reduce body sway", *Neuroscience Letters*, vol. 722, 134799, 2020, doi: 10.1016/j.neulet.2020.134799

[13] Hessels, R.S. and I.T.C. Hooge, "Eye tracking in developmental cognitive neuroscience - The good, the bad and the ugly", *Developmental Cognitive Neuroscience*, vol. 40, 100710, 2019, doi: 10.1016/j.dcn.2019.100710

[14] Juhola, M., H. Aalto, and T. Hirvonen, "Using results of eye movement signal analysis in the neural network recognition of otoneurological patients", *Computer Methods and Programs in Biomedicine*, vol. 86, no.3, pp. 216-226, 2017, doi: 10.1016/j.cmpb.2007.02.008

[15] Juhola, M., T. Tossavainen, and H. Aalto, "Influence of lossy compression on eye movement signals", *Computers in Biology and Medicine*, vol. 34, no.3, pp.221-239, 2004, doi: 10.1016/S0010-4825(03)00059-3

[16] Enkelejda Kasneci, Thomas Kübler, Klaus Broelemann, et al., "Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth", *Computers in Human Behavior*, vol. 68, pp. 450-455, 2017, doi: 10.1016/j.chb.2016.11.067

[17] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, et al., "Deep learning for healthcare applications based on physiological signals: A review". *Computer Methods and Programs in Biomedicine*, vol. 161, pp.1-13, 2018, doi: 10.1016/j.cmpb.2018.04.005

[18] Jue Li, Heng Li, Waleed Umer, et al., "Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology", *Automation in Construction*, vol.109, 103000, 2020, doi: 10.1016/j.autcon.2019.103000

[19] Xue Wang, Lin Lin, Meiqi Han, et al., "Impacts of cues on learning: Using eye-tracking technologies to examine the functions and designs of added cues in short instructional videos", *Computers in Human Behavior*, vol. 107, 106279, 2020, doi: 10.1016/j.chb.2020.106279

[20] Dexiang Zhang, Jukka Hyönä, Lei Cui, et al., "Effects of task instructions and topic signaling on text processing among adult readers with different reading styles: An eye-tracking study", *Learning and Instruction*, vol. 64, 101246, 2019, doi: 10.1016/j.learninstruc.2019.101246

[21] Król, M.E. and M. Król, "A novel machine learning analysis of eye-tracking data reveals suboptimal visual information extraction from facial stimuli in individuals with autism", *Neuropsychologia*, vol.129, pp.397-406, 2019, doi: 10.1016/j.neuropsychologia.2019.04.022

[22] Angelina Vernetti, Atsushi Senju, Tony Charman, "Simulating interaction: Using gaze-contingent eye-tracking to measure the reward value of social signals in toddlers with and without autism", *Developmental Cognitive Neuroscience*, vol. 29, pp.21-29, 2018, doi: 10.1016/j.dcn.2017.08.004

[23] Vettori, S., et al., "Combined frequency-tagging EEG and eye tracking reveal reduced social bias in boys with autism spectrum disorder", *Cortex*, vol.125, pp.135-148, 2020, doi: 10.1016/j.cortex.2019.12.013

[24] Ding, X., & Lv, Z., "Design and development of an EOG-based simplified Chinese eye-writing system", *Biomedical Signal Processing and Control*, vol. 57, 101767, 2020, doi: 10.1016/j.bspc.2019.101767

[25] Dua, M., Gupta, R., Khari, M., et al., "Biometric iris recognition using radial basis function neural network", *Soft Computing*, vol. 23, no. 22, pp.11801-11815, 2019, doi: 10.1007/s00500-018-03731-4

[26] Dutta, A., Mondal, A., Dey, N., et al., "Vision Tracking: A Survey of the State-of-the-Art", *SN Computer Science*, vol.1, no.1, pp. 57, 2020, doi:10.1007/s42979-019-0059-z

[27] Khari, M., Garg, A. K., Crespo, R. G., et al., "Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks", *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 5, no. 7, pp.22-27, 2019, doi: 10.9781/ijimai.2019.09.002

[28] Ahuja, R., Jain, D., Sachdeva, D., et al., "Convolutional Neural Network

Based American Sign Language Static Hand Gesture Recognition", *International Journal of Ambient Computing and Intelligence*, vol. 10, no.3, pp. 60-73, 2019, doi: 10.4018/IJACI.2019070104

[29] Deng, M., Meng, T., Cao, J., et al., "Heart sound classification based on improved MFCC features and convolutional recurrent neural networks", *Neural Networks*, vol.130, pp. 22-32, 2020, doi: 10.1016/j.neunet.2020.06.015

[30] Raj, R., Rajiv, P., Kumar, P., et al., "Feature based video stabilization based on boosted HAAR Cascade and representative point matching algorithm", *Image and Vision Computing*, vol. 101, 103957, 2020, doi: 10.1016/j.imavis.2020.103957

[31] Dey, N., Ashour, A. S., & Hassanien, A. E., "Feature detectors and descriptors generations with numerous images and video applications: a recap", *Feature detectors and motion detection in video processing*", 2017, pp. 36-65, IGI Global.

[32] Zhou, X., et al., "Eye tracking data guided feature selection for image classification", *Pattern Recognition*, vol. 63, pp. 56-70, 2017, doi: 10.1016/j.patcog.2016.09.007

[33] Ali, M. N. Y., Sarowar, M. G., Rahman, M. L., et al., "Adam deep learning with SOM for human sentiment classification", *International Journal of Ambient Computing and Intelligence*, vol. 10, no. 3, pp. 92-116, 2019, doi: 10.4018/IJACI.2019070106

[34] Shah, S., Kumar, A., Kumar, R., & Dey, N., "A robust framework for optimum feature extraction and recognition of P300 from raw EEG", *U-Healthcare Monitoring Systems*, 2019, pp. 15-35, Academic Press.

[35] Dey, A., Bhattacharya, D. K., Tibarewala, D. N., et al., "Chinese-chi and Kundalini yoga Meditations Effects on the Autonomic Nervous System: Comparative Study", *International Journal of Interactive* Multimedia & Artificial Intelligence, vol. 3, no. 7, pp. 87-95, 2016, doi:10.9781/ijimai.2016.3713

[36] Wu, H., & Zhao, X., "A Small and Portable Foot Motion Recognition Device Used in VR Environment", *International Journal of Ambient Computing and Intelligence*, vol. 10, no.3, pp. 1-16, 2019, doi: 10.4018/ijaci.2019070101

[37] Gopal, B., & Manohar, S., "VLSI architecture for the Winograd Fourier transform algorithm", *Microprocessing and Microprogramming*, vol. 40, no. 9, pp. 605-616, 1994, doi: 10.1016/0165-6074(94)90089-2

[38] Prabhakar, D. V. N., Sreenivasa Kumar, M., & Gopala Krishna, A., "A Novel Hybrid Transform approach with integration of Fast Fourier, Discrete Wavelet and Discrete Shearlet Transforms for prediction of surface roughness on machined surfaces", *Measurement*, vol. 164, 108011, 2020, doi: 10.1016/j.measurement.2020.108011

[39] George, A. E. W., So, S., Ghosh, R., & Paliwal, K. K., "Robustness metric-based tuning of the augmented Kalman filter for the enhancement of speech corrupted with coloured noise", *Speech Communication*, vol. 105, 62-76, 2018, doi: 10.1016/j.specom.2018.10.002

[40] Olbrys, J., & Mursztyn, M., "Estimation of intraday stock market resiliency: Short-Time Fourier Transform approach", *Physica A: Statistical Mechanics and its Applications*, vol. 535, 122413, 2019, doi:

[41] Mateo, C., & Talavera, J. A., "Short-Time Fourier Transform with the Window Size Fixed in the Frequency Domain (STFT-FD): Implementation", *SoftwareX*, vol. 8, pp.5-8, 2018, doi: 10.1016/j.physa.2019.122413

[42] Y. LeCun, L. Bottou, Y. Bengio et al., "Gradient-Based Learning Applied to Document Recognition", *Proceedings of the IEEE*, vol. 86, no.11, pp.2278-2324, 1998, doi: 10.1109/5.726791

[43] Hande Erkaymaz, Mahmut Ozer, İlhami Muharrem Orak, "Detection of directional eye movements based on the electrooculogram signals through an artificial neural network", *Chaos, Solitons & Fractals*, vol. 77, pp. 225-229, 2015, doi: 10.1016/j.chaos.2015.05.033

[44] Bo Liu, Yanshan Xiao, Longbing Cao, "SVM-based multi-state-mapping approach for multi-class classification", *Knowledge-Based Systems*, vol. 129, pp.79-96, 2017, doi: 10.1016/j.knosys.2017.05.011

[45] Ying Wang, Qun Wu, Nilanjan Dey, et al., "Deep back propagation–long short-term memory network based upper-limb sEMG signal classification for automated rehabilitation", *Biocybernetics and Biomedical Engineering*, vol. 40, no.3, pp. 987-1001, 2020, doi: 10.1016/j.bbe.2020.05.003

[46] Yu Wang, Yating Chen, Ningning Yang, et al., "Classification of mice hepatic granuloma microscopic images based on a deep convolutional neural network", *Applied Soft Computing*, vol. 74, pp. 40-50, 2019, doi: 10.1016/j.asoc.2018.10.006

[47] Dan Wang, Zairan Li, Nilanjan Dey, "Optical pressure sensors based plantar image segmenting using an improved fully convolutional network", *Optik*, vol. 179, 99-114, 2019, doi: 10.1016/j.ijleo.2018.10.155

[48] Zhenzhen Song, Longsheng Fu, Jingzhu Wu, et al., "Kiwifruit detection in field images using Faster R-CNN with VGG16", *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 76-81, 2019, doi: 10.1016/j.ifacol.2019.12.500

[49] Pengjie Tang, Hanli Wang, Sam Kwong, "G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition", Neurocomputing, vol. 225, pp. 188-197, 2017, doi: 10.1016/j.neucom.2016.11.023

[50] Ashkan Shakarami, Hadis Tarrah, AliMahdavi-Hormat, "A CAD system for diagnosing Alzheimer's disease using 2D slices and an improved AlexNet-SVM method", *Optik*, vol. 212, 164237, 2020, doi: 10.1016/j.ijleo.2020.164237

[51] M. G. Huddar, S. S. Sannakki, V. S. Rajpurohit, "Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, 2020, doi:10.9781/ijimai.2020.07.004

Yuanfeng Li

He received his B.Eng. degree from Zhejiang A&F University in 2012, Master of Design Science in 2015 from Zhejiang Sci-Tech University. In 2016, he was appointed as a Lecturer at Jiyang College of Zhejiang A&F University. His research interests include human factors engineering and human-computer interaction focusing on user experience designs.

Jiangang Deng

He received Bachelor of Arts from Zhejiang A&F University in 2003, Master of Agriculture from Beijing Forestry University in 2009. In 2008, he has appointed as a Lecturer at Zhejiang A&F University. His topic research is product innovation design, human interactive design, intelligent design, green design and computer aided conceptual design.

Qun Wu

Dr. Wu is an Associate Professor of Human Factor at the Institute of Universal Design, Zhejiang Sci-Tech University, China. He received his Ph.D. in College of Computer Science and Technology from Zhejiang University, China, in 2008. He holds a B.E. degree in Industrial Design from Nanchang University, China, in 2001, and a M.E. degree in Mechanical Engineering from Shaanxi University of Science and Technology, China, in 2004. His research interests include machine learning, human factor and product innovation design.

Ying Wang

She received the B.Eng. degree in Digital Art and Design from Zhejiang University, China in 2014. She is now employed as an Associate. Professor in Department of Industrial Design at College of Art and Design, Zhejiang Sci-tech University, Hangzhou, China. Her research topic is universal design, HCI for elderly, biosensor based technologies for interactive design, rehabilitation engineering design, and computer aided intelligent design.

Deep Multi-Model Fusion for Human Activity Recognition Using Evolutionary Algorithms

Kamal Kant Verma[1]*, Brij Mohan Singh[2]

[1] Research Scholar, Department of CSE, Uttarakhand Technical University, Dehradun (India)
[2] Department of Computer Science and Engineering College of Engineering Roorkee, Roorkee (India)

Received 30 October 2020 | Accepted 14 May 2021 | Published 6 August 2021

Abstract

Machine recognition of the human activities is an active research area in computer vision. In previous study, either one or two types of modalities have been used to handle this task. However, the grouping of maximum information improves the recognition accuracy of human activities. Therefore, this paper proposes an automatic human activity recognition system through deep fusion of multi-streams along with decision-level score optimization using evolutionary algorithms on RGB, depth maps and 3d skeleton joint information. Our proposed approach works in three phases, 1) space-time activity learning using two 3D Convolutional Neural Network (3DCNN) and a Long Sort Term Memory (LSTM) network from RGB, Depth and skeleton joint positions 2) Training of SVM using the activities learned from previous phase for each model and score generation using trained SVM 3) Score fusion and optimization using two Evolutionary algorithm such as Genetic algorithm (GA) and Particle Swarm Optimization (PSO) algorithm. The proposed approach is validated on two 3D challenging datasets, MSRDailyActivity3D and UTKinectAction3D. Experiments on these two datasets achieved 85.94% and 96.5% accuracies, respectively. The experimental results show the usefulness of the proposed representation. Furthermore, the fusion of different modalities improves recognition accuracies rather than using one or two types of information and obtains the state-of-art results.

Keywords

Human Activity Recognition, Support Vector Machine, 3D-Convolutional Neural Network, LSTM, Deep Learning, Genetic Algorithm, Particle Swarm Optimization.

DOI: 10.9781/ijimai.2021.08.008

I. Introduction

Human activity recognition (HAR) is a demanding research topic from last two decades in computer vision. HAR's objective is to learn dynamically and automatically of human activities such as drinking, eating, clapping, walking, etc. Human activities are mainly associated with daily human activities, indoor and outdoor activities. HAR has many applications in multiple domains, such as human security applications, health-care sectors, virtual reality games, intelligent monitoring systems, and human-computer applications [1]-[2]. Moreover, activity recognition is a challenging task, due to the structural changes among the subjects, inter class and intra class similarity between the activities. Additionally, some issues still exist, such as cluttered background, occlusion, camera motion, multi-camera recognition, complex scenes, human to human interaction, or human to object interaction that makes video-based human activity recognition systems more complex and challenging [3]. Systematically analyzing and recognition of human postures can make people understand the sense of human behavior. This could be advantageous to augment the monitoring process of indoor activities, understand normal and abnormal activities, and to maintain security surveillance systems in real life. Hence, the human activities recognition system has become a vital issue nowadays in both academia and industry.

* Corresponding author.

E-mail address: kkv.verma@gmail.com

From the past few years, there is a significant improvement in this research area, due to two main factors. Firstly, frequent accessibility of low-cost depth camera, secondly the robust features learning using deep convolutional neural network. Before the advent of the RGB-D (Depth sensor), the research on activity recognition was limited to the recognition of human activities from frame sequences captured using traditional RGB video cameras [3]. As the depth camera like Microsoft Kinect, Xbox 360 Kinect, and ASUS Xtion came into existence, the research has shifted towards the learning of human activities using depth information [4]-[6].

Unlike the RGB, the depth camera has several advantages, for example, there is no effect of variable lighting conditions, illumination changes. Segmentation of shape and structure given by the depth camera are easy to handle, insensitive to the cluttered background. It also gives additional modality like 3D skeleton joint positions that deliver multi-dimensional data for better activity recognition. Zhu et al. [7] proposed a robust feature-fusion approach using 3D-skeleton joints information to capture spatial-temporal features.

Currently, human activity recognition techniques can be classified into two groups: traditionally handcrafted feature extraction techniques and using deep learning techniques. In the previous study, the traditional handcrafted features extraction methods include Histogram of Oriented Gradients (HOG) and Histogram of optical flow (HOF) from 2D-images [9]. In a recently published article, Franco et al. [10] used RGB video sequences and skeleton joint positions for activity recognition. They used HOG to learn motion information

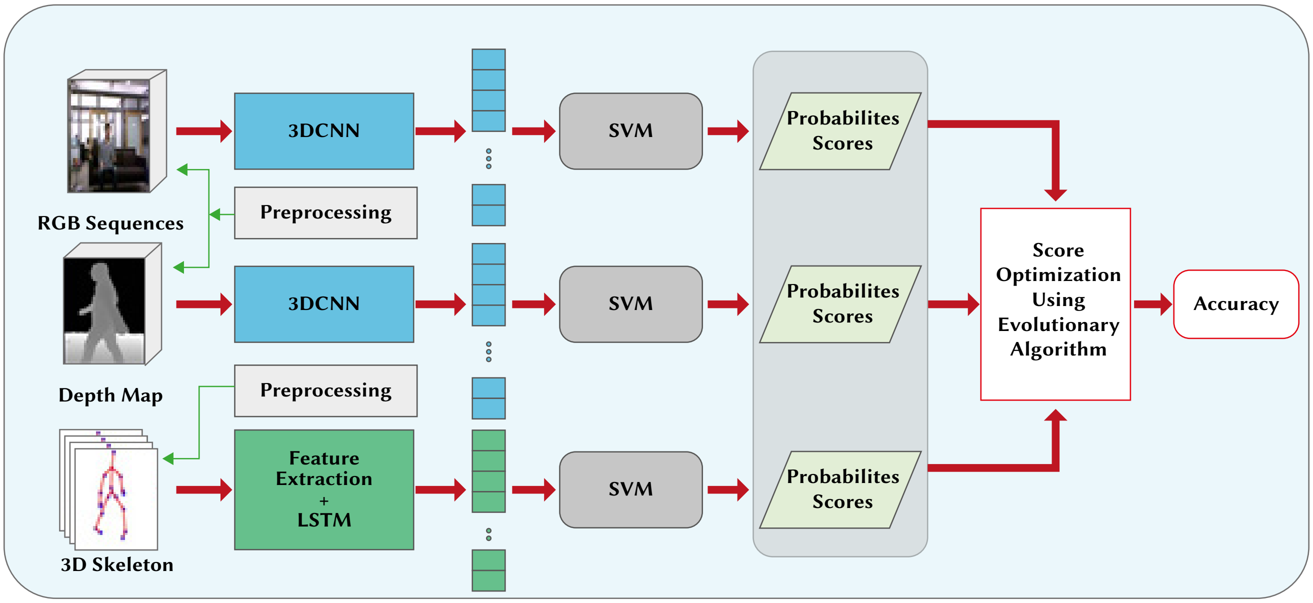


Fig. 1. Flow diagram of the proposed system.

from RGB video sequences and bag-of-words (BOW) for spatial information learning from the skeleton postures. However, the skeleton joints position information extracted from the depth map contains several difficulties like occlusion due to any arbitrary object present in the scene, lack of precision, and viewpoint variation [11]. Therefore, to prepare a robust action recognition system, these errors can be eliminated by finding the body pose estimation component. Hence, Chaaraoui et al. [8] proposed a method that combines body pose estimation and 2D shape clue parameters to make a robust action recognition system. They extracted the low dimensional features from skeleton joints, fetched silhouette from depth images, and combined both to improve the activity recognition. Pham et al. [12] and Yang et al. [13] also used in-depth information about the 3D structure of human body parts and body posture information.

Similarly, Jalal et al. [14] also used human silhouette as additional information from depth video. They used R-transformation to find the translational and scaling view-invariant features, the principal component analysis (PCA) is applied to make the features low dimensional. Furthermore, they also used linear discriminant analysis (LDA) to obtain prominent features, and finally, the Hidden Markov Model (HMM) is used to train and identify human activities.

However, the methods discussed in the above section only utilize handcrafted features. Though the extractions of the handcrafted features are a challenging task, and they do not require a large amount of training data. Moreover, it is a time-consuming process also, and these are the standard methods and are not ideal for a specific task such as activity recognition from complex scenes. Therefore handcrafted features may not thoroughly learn the intrinsic spatial and temporal information present in RGB-D sequences thus, these methods could not satisfy as per desire.

From the past few years, deep neural networks have gained considerable achievements in computer vision using deep architectures. A deep neural network is an ordered arrangement of multiple layers that learns the features automatically. In DNN, each layer gets the output as an input from the previous layer and performs nonlinear transformation. One such deep neural network is convolutional neural network (CNN) that gets better recognition results. In the era of computer vision, CNN achieves enough popularity in vision based applications such as surveillance [15], image classification [16], image segmentation [17], object classification, medical sector [18], object recognition [19], recognition of facial expression [20], biometric and agriculture [21], video classification [22] and many others [23]. CNN is enough capable to handle a large amount of data with a high level of accuracy [18].

However, the trainable feature extraction using CNN called CNN-features and recognition of CNN-features using SVM (CNN-features + SVM) architecture has achieved remarkable advantage in the previous literature. For example, Niu et al. [24] evaluated CNN-features + SVM architecture on 10000 grayscale images of MNIST dataset, and showed that the CNN-features + SVM combination generated highest 99.81% recognition accuracy. Similarly, Xue et al. [25] have proposed NBI-Net architecture for cancer detection which is again based on CNN as trainable feature extractor and SVM as predictor. The proposed architecture was trained and tested on 6500 patch samples and achieved 90.93% recognition accuracy which was 0.9% more as compare if only CNN had been used alone for recognition. Likewise, Sargano et al. [26] have also used CNN-SVM architecture for human activity recognition. They used pre-trained CNN network as trainable feature extractor and SVM for recognition of human activities. The proposed architecture has been evaluated on two UCF-Sports and KTH datasets and achieved highest accuracy over mentioned state-of-the-art methods. Inspired by the above discussed outstanding performance we have also used CNN-features + SVM architecture to process RGB and depth data. Another reason for considering CNN-SVM architecture is because both CNN and SVM have already shown excellent performance on human activity recognition independently. Therefore, in this work we have focused on their fusion to come out their best qualities. On the other hand, RNNs networks have also shown the exceptional performance on sequence modeling applications such as image captioning [27], video analysis [28], and language translation [29]. Due to this reason, this work used RNN especially LSTM network to process skeleton data.

Likewise, deep multi-model approaches [30]-[32] have substantially outperformed over the manual features extraction methods [11]-[14]. Therefore, this paper also suggests a multi-model approach which uses CNN-features + SVM architecture to process RGB, depth data and a LSTM network to process skeleton joint positions. The SVM classifier with RBF kernel has been used corresponding to classification using features extracted in each parallel channels for RGB, depth and skeleton data. The flow diagram of our proposed approach is given in Fig. 1. The proposed approach uses three different types of neural network: two 3DCNN for RGB and depth videos and a LSTM network for skeleton stream.

- First, spatio-temporal feature learning has been implemented from RGB and Depth video sequences using two different 3DCNN models. At the same time, a LSTM network has been trained using a feature vector obtained from the skeleton data.
- Secondly, in order to find the class score of each test activity from individual streams, three SVM networks have been trained using the spatial-temporal features that have been extracted using three different deep neural networks for each modality.
- At last, an evolutionary algorithm is used to optimize the class score obtained for the three SVM networks during score-level fusion from the individual streams to find the optimized class label of each test activity.
- We have trained and tested our proposed method on MSRDailyActivity3D and UTKinectAction3D Datasets.

The remaining sections of the paper are organized as: The relevant literature review is given in the section II. Section III contains our proposed work for activity recognition. Section IV consists of experimental work, results and discussions. Section V contains conclusion followed by future work.

II. Related Work

Activity recognition using RGB-D is a hot research area from past several years. Hence, in this section we covered previous literature using RGB, Depth videos, skeleton joints and hybrid data.

A. Activity Recognition Using RGB Data Only

Activity recognition in RGB videos is a most frequent task because of availability of RGB video activity datasets, for example Soomro et al. [33] developed the UCF-101 dataset containing 101 different activity classes with more than 13K video clips. Kuehne et al. [34] developed another commonly used dataset HMDB-51 containing 51 activity categories with 7000 video clips. In addition to that, Caba Heilbron et al. [35] proposed the large scale activity dataset ActivityNet. The ActivityNet dataset contains 203 different activity classes with approximately 137 untrimmed video clips per class. Besides these datasets, several others RGB datasets also exist for activity recognition task [36]. Prior to the learning based approaches, numerous handcrafted-based feature descriptors methods have been suggested by the researchers in past literature [36]. A lot of handcrafted-features based approaches are presented in [36], that are ranges from space-time features descriptors, appearance-based (shape and motion) and fuzzy logic based approaches, etc. Out of them dense trajectory [37] methods have achieved best results compare to others.

Later, the convolutional neural network has been efficiently used for activity recognition in RGB video classification due to the remarkable performance of deep neural network (DNN) in image and video classification [38]. Simonyan et al. [39] proposed a two-stream action recognition approach for spatial and temporal information. The spatial information is extracted from the frame using spatial stream convolutional neural network (ConvNet) and temporal information is extracted from the frames using optical flow displacement field vectors (OFDF). Both the streams are combined together for late fusion. The average fusion and multi-labels SVM are used for classification. The method is tested on UCF-101 and HMDB-51 datasets and obtained 88.0% and 59.4% accuracy respectively. Karpathy et al. [40] proposed multi-resolution CNN with two streams. The first stream is fovea stream and second stream is context stream. Both the streams take the consecutive frames as an input. The extracted information from both the streams are fused using different fusion techniques such as early, late and slow fusion. The proposed approach is validated on large scale activity dataset named sports 1M that contains over one millions sports videos from youtube videos which is divided into 487 different classes. Tran et al. [41] used 3DCNN to learn spatio-temporal features from the video sequences simultaneously. They applied varying kernel size and found that 3×3×3 kernel outperforms among the used kernel size. The proposed approach in [41] produces superior results on four datasets: UCF-101, ASLAN, YUPENN and UMD datasets. Feichtenhofer et al. [42] suggested the features-level fusion strategy to take the benefit of spatio-temporal information. Spatial-temporal features fusion at the last convolution layer are more advantageous rather than fusing at softmax layer and also decreases the parameter counts. The proposed approach is implemented on two benchmark datasets UCF-101 and HMDB-51 and gave 93.5% and 69.2% accuracies. Varol et al. [43] have given a convolutional neural network based on long-term temporal information (LTC-CNN). They implemented LTC-CNN with varying number of temporal information ranges from t=16 to t=100 and found that as the temporal information t increases, the value of the accuracy also increases. They experimented their approach on two challenging datasets, UCF-101 and HMDB-51, and achieved 92.7% and 67.2% accuracies respectively. Ullah et al. [44] gave a novel framework for activity recognition using CNN and deep-BLSTM. In this framework deep features are extracted from the every sixth frame of a video sequence using CNN. Next, the features learning have been performed by deep-BLSTM network. This approach is tested on UCF-101, Youtube-11 and HMDB-51 datasets and obtained 91.21%, 92.84% and 87.64 accuracies comparable to the state-of-the-art results. Recently Verma et al. [45] used powerful coarse and file level classification framework for single and multi-limb activity recognition. The advantage of this framework is that, it divides the losses at multiple levels. They implemented their approach on UTKinectAction3D dataset and achieved 99.92% accuracy at coarse level with overall 97.88% accuracy.

For the spatial-temporal activities learning from the RGB information, most of the above discussed methods used deep convolutional neural networks. Unlike the above methods, the proposed approach uses deep convolution neural networks for feature learning with support vector machines for score prediction from RGB streams.

B. Activity Recognition Using Depth Data Only

With the advent of depth cameras such as Kinect, Xbox 360, and ASUS Xtion, depth video sequence drastically changed the research pattern. It has several advantages such as it is constant with respect to the background changes and contains depth information which is not present in RGB. Recently, Li et al. [46] proposed an idea of mapping the frames onto three orthogonal planes top view, side view and front view, then they formed the depth motion map (DMM) by stacking all three images. Local ternary pattern are applied to DMM to filter out the images and finally they used CNN for classification. The method in [46] gained state-of-the-art accuracy on MSRAction3D and MSRGesture3D datasets and achieved 98.81% and 99.67% accuracies respectively. Wang et al. [47] proposed Dynamic Depth Image (DDI), Dynamic Depth normal image (DDNI) and Dynamic Depth Motion Normal Image (DDMNI) form the depth videos. DDI is used to capture motion dynamics of the sequence while the structural information of a posture is extracted using both DDNI and DDMNI. At last, convolutional neural network (ConvNet) is used for activity classification. The proposed method is implemented on Large-scale Continuous Gesture Recognition, Large-scale Isolated Gesture Recognition and NTU RGB+D datasets, and achieved 59.21%, 87.08% and 84.22% cross view accuracies which is greater than state-of-the-art accuracies. In addition to that, Chen et al. [48] also presented an effective depth motion map based local binary pattern (DMMs-LBP) framework for activity recognition. In this approach three motion maps have been generated using frame differencing after mapping the depth images onto three orthogonal planes related to top view, side view and front view. Then, a local binary pattern (LBP) descriptor has been used to find the LBP histogram for each DMM and represented by a feature vector. The method suggested in [48] is implemented on MSRAction3D and MSRGesture3D datasets and obtained 87.9% and 96.4% accuracies respectively. Similarly, Wang et al. [49] also proposed a framework known as hierarchal depth motion map (HDMM) with three channel convolutional neural networks related to front, side and top view. The method is tested on MSRAction3D, MSRAction3DExt, UTKinectAction3D and MSRDailyActivity3D datasets and achieved state-of-the-art results. However, Megavannan et al. [50] used handcrafted features prior to the popularity of deep neural network. They used motion history images (MHI), average depth image (ADI) and depth difference images (DDI) in order to find the space-time features. Then two features have been generated, firstly Hu movement feature is calculated using MHI and ADI and in the second case DDI is used to find the hierarchically division. They also created their own dataset having eight activities to validate their approach and obtained overall 90% accuracy.

Even though most of the above methods used depth motion map (DMM) to learn temporal information, our spatial-temporal feature learning policy for depth channel utilizes 3D deep convolutional neural network (3DCNN) and further find the class score using conventional support vector machine. This combination of 3DCNN+SVM outperforms over traditional handcrafted features and DMM strategy.

C. Activity Recognition Using Skeleton Data Only

Besides the limitation of the depth maps, the skeleton joint positions have an advantage that it is three dimensional in nature compared to the one dimensional depth map and two dimensional RGB data. Furthermore, skeleton positions are also considered as temporal information, therefore many past approaches have attempted to use RNN network specially LSTM. Recently, Han et al. [51] proposed global spatio-temporal (GL-LSTM) model for activity recognition. The GL-LSTM architecture combines the accumulative learning curve (ALC) for temporal information and global spatial attention (GSA) to prepare the spatial information. Then LSTM network with difference clue has been used for action classification. This approach is validated on NTU RGB-D and SBU datasets and generated results outperform the given state-of-the-art methods. Similarly, Ren et al. [52] also discussed several action recognition methods using deep recurrent neural network (DRNN), convolutional neural network (CNN), and graph convolutional neural network (GCNN). They also discussed latest skeleton datasets with their performance. Another skeleton based coarse and fine level continuous human activity recognition framework is suggested by Saini et al. [53]. In this, all activities have been segmented into two categories such as sitting and standing. Next, five features such as 3D-joints, angular movement, angular direction, distance and velocity have been extracted for activity classification using BLSTM network. For evaluation of the approach, they captured 1110 continuous activity sequences for 24 activity classes using Kinect depth sensors and achieved 68.9% and 64.45 % accuracies through length modeling and without length modeling respectively. Chikhaoui et al. [54] proposed a 3D skeleton based aggressive behavior learning framework, this framework is based on the fusion of two features such as joint based and body part based in order to learn the spatio-temporal information. Then the combined feature vector is used by ensemble learning based rotation forest. The proposed approach is tested on various 3D activity datasets for instance TRI, Kintense, UTKinectAction3D, Florence Action and MSRAction3D datasets and absolutely distinguishes the various activity categories for each dataset. Shahroudy et al. [55] suggested a novel part aware LSTM (P-LSTM) in which local structure of five different body parts (two legs, two hands and torso) are individually mapped. These five different memory cells are combined to find the global information. They have evaluated their work on a self-developed NTURGB-D dataset and have shown the effectiveness of the proposed P-LSTM network on the traditional recurrent neural network.

The above discussion exhibits the better outcome of DRNN specially LSTM for activity recognition using 3D skeleton joint positions. It also shows that the LSTM network has improved performance over other deep recurrent neural networks for sequence learning problem in 3D domain. Therefore, we have also used LSTM network to learn space-time features from skeleton data.

D. Action Recognition Using Hybrid (RGB+Depth+Skeleton) Data

In past literature, combining the RGB, depth and skeleton joints information has given promising results for human activity recognition. Recently, Gu et al. [56] proposed a novel framework which combines domain knowledge clue parameter for decision making. Three motion history images (MHIs) have been used to learn global information. Local spatial and temporal information are extracted using skeleton joint positions, and this information is fused together with domain knowledge clue parameter. The proposed work is evaluated on two RGB-D datasets and has given the best results on the mentioned stated-of-the-art methods. Khaire et al. [57] proposed a 5-stream CNN for activity recognition using MHI, DMM and skeleton images. Moreover, this work has given a new way of creating skeleton images. The given approach is implemented on CAD-60, SUB Kinect interaction and UTD-MHAD datasets and got comparative results on the state-of-the-arts methods. Tomas et al. [58] presented a method that uses CNN and stacked Auto Encoder (SAE). CNN is used to learn motion information from motion history images and SAE is used to find the physical structural information of a static posture. The proposed method is tested on two benchmark datasets MSRDailyActivity3D and MSRAction3D and achieved 91.3% and 74.6% accuracy respectively. Ijjina et al. [59] proposed an approach that utilizes the key pose related to each action. They captured motion information using two temporal MEI and MHI. Then they took the frame differencing from each category. The temporal information using MEI and MHI have been captured from RGB and depth video sequences and placed as an input to the CNN for action recognition. The suggested method is implemented on four RGB-D datasets namely, MIVIA action, Weizmann, SBU Kinect action, and NATOPS gesture datasets and achieved 93.37%, 100%, 90.98% and 86.58% accuracies. Zhao et al. [60] also proposed a multi-model approach for human behavior recognition using RGB, depth and skeleton data. They computed space-time features from each modality using deep 3DCNN and obtain class probability scores of each test activity using SVM. Next, achieved class probability scores of each test activity are fused together using weighted linear combination techniques. Subsequently, they implemented their approach on two RGB-D datasets such as MSRAction3D and MSRDailyActivity3D, and achieved 94.15% and 97.29% accuracies.

After the above discussion, some of the approaches use all three modalities like RGB, depth and skeleton joints positions, while some use RGB and skeleton, depth and skeleton while some others use RGB and depth sequences. Therefore, next, in the section III, we have discussed our Deep Multi-Model (DMM) fusion approach which combines RGB, depth sequence and skeleton joints positions for human activity recognition system.

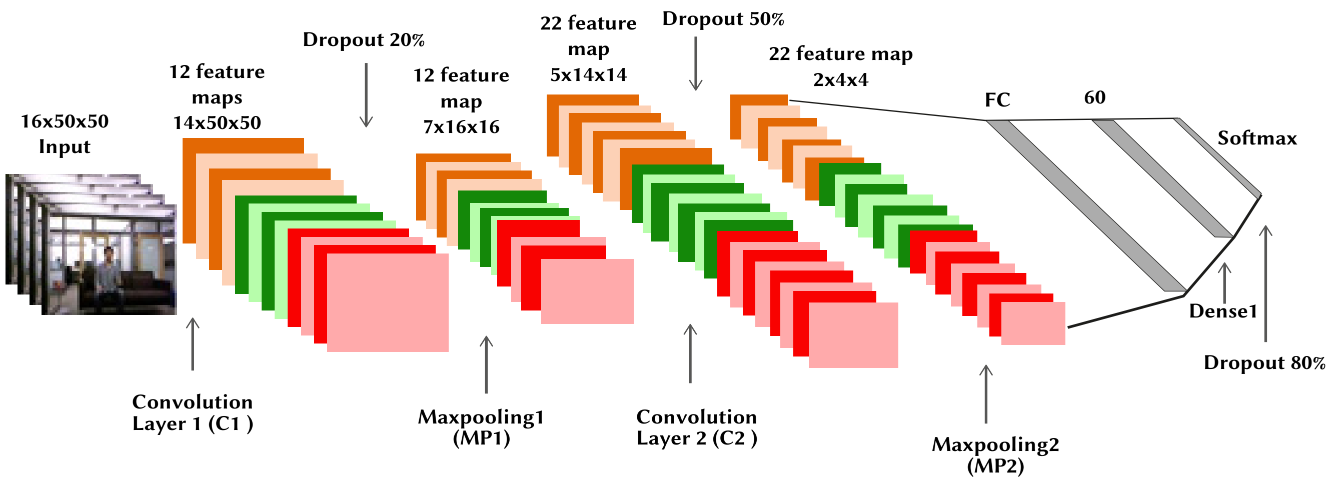


Fig. 2. Architecture of 3DCNN for RGB video sequences.

III. Proposed Work

In the proposed work to recognize the human activity, a novel multi-model approach with different modalities –RGB, Depth and skeleton joint information –using evolutionary algorithms has been given. The proposed work is done in three different levels. In the first level, spatial-temporal features have been extracted from different modalities, in the second level three independent SVMs are trained using the features extracted from the first level, and in the last level the probability scores are fused and optimized using two evolutionary algorithms such as GA and PSO.

The spatial-temporal activity learning from each modality has been discussed in the subsequent sections.

A. Spatio-Temporal Activity Learning From RGB Information Using 3DCNN

The 3D Convolutional neural network (3DCNN) is a deep neural network introduced in [61], used to learn the features from both spatial and temporal dimension. The 3D convolutional is accomplished by convolving a 3-dimensioanl filter over the cube obtained by stacking the spatial and temporal frames one after another. In order to learn the motion related information from the sequence of frames, features maps exist in the convolution layer that are connected with the multiple contiguous frames from the previous layer. Multiple convolutional layers are used to obtain both lower and higher level features. Hence the design technique of Convolution neural network is the increase the feature maps by increasing the number of layers in the network. Therefore, a 3D Convolutional neural network is obtained by convolving the 3D filter kernel over the multiple staked fame to produce a 3D cube. Finally, the value in the jth feature map of ith layer at a position (x, y, z) is given in equation (1).

$$v_{i,j}^{x,y,z} = \tanh\left(b_{ij} + \sum_m \sum_{a=0}^{A_{i-1}} \sum_{b=0}^{B_{i-1}} \sum_{c=0}^{C_{i-1}} w_{ijm}^{abc}\, v_{(i-1)m}^{(x+a)(y+b)(z+c)}\right) \quad (1)$$

Where m is the index value of the feature maps in the (i-1)th layer, which are connected to the present feature map and b_{ij} is the bias value of the mth feature map. The function tanh () is the hyperbolic tangent function. The values A_i and, B_i represent the height and width of the 3D-kernel and C_i is the dimension of the 3D-kernel along the temporal direction. The term w_{ijm}^{abc} is the (a, b, c)th value of the kernel of the mth feature map in the previous layer.

In order to learn the spatio-temporal features from the RGB video sequences, a 3DCNN is used. The proposed 3DCNN contains two convolution layers and two max-pooling layers. To prepare the input for the 3DCNN network, simple preprocessing has been performed in which each frame is resized to a fixed height and width dimension. The height and width is resized to [50, 50] using the Open CV library in python. Since each video sequence has a varying number of frame length so it is difficult to normalize all video sequences to the same length, hence a fixed input sequence is used as a depth dimension. The value of the input depth dimension was set to 16. Therefore the size of the RGB input cube which is given as an input to the 3DCNN is [16(depth) ×50(height) ×50(width)]. This input cube is processed by the first convolution layer (C1) followed by the first maxpooling layer (ML1). A dropout layer with an amount of 20% is used after the first convolution layer. The network also contains a second convolution layer (C2) followed by a second maxpooling layer (MP2). A 50% dropout is also used between the C2 and MP2. Then a fully connected layer (FC) layer is used to obtain the one dimensional feature vector. Then, a dense layer with 60 neurons is used. Finally a softmax layer is used for classification followed by a dropout layer with 80% dropout amount. The number of feature maps in C1 and C2 are 12 and 22 respectively. The size of the kernel in first and second convolution layers are (3×1×1) and (3×3×3) respectively. For the first and second 3D maxpooling layers, (2×3×3) down sampling is used for both MP1 and MP2 layers. We have used the same network parameter to process RGB video sequences for both MSRDailyActivity3D and UTKinectAction3D datasets. The architecture of the 3DCNN for spatio-temporal features learning from RGB video sequences is shown in Fig. 2.

B. Spatio-Temporal Activity Learning From Depth Information Using 3DCNN

Similarly, for the spatio-temporal feature learning from the depth map sequence, a different 3DCNN is used with different network parameters. During the preprocessing step, the input cube is formed in the similar way as we performed in RGB, except that different length, width and depth dimension are applied. The size of the input depth cube which is inserted to the 3DCNN is [13(width) ×32(height) ×32(width)]. The input depth cube is processed by first 3D convolution layer (DC1) followed by maxpooling layer (DMP1). A dropout of 20% has been used in between. Another 3D convolutional layer (DC1) is used followed by a second (DMP2) layer. Similarly, 20% dropout is also used between the second 3D convolutional layer and second maxpooling layer DC1 and DMP2. Then, a fully connected (FC) layer is used to find the one-dimensional feature vector, followed by a dense layer with 128 neuron units. At last a softmax layer is used for classification. We have also used a dropout layer with 50% amount before the softmax layer. The count of feature maps for DC1 and DC2 layers are 16 and 32. The size of the kernel used in DC1 and DC2 layers are (3×3×3) and (5×5×5) respectively. The downsampling size of 3D

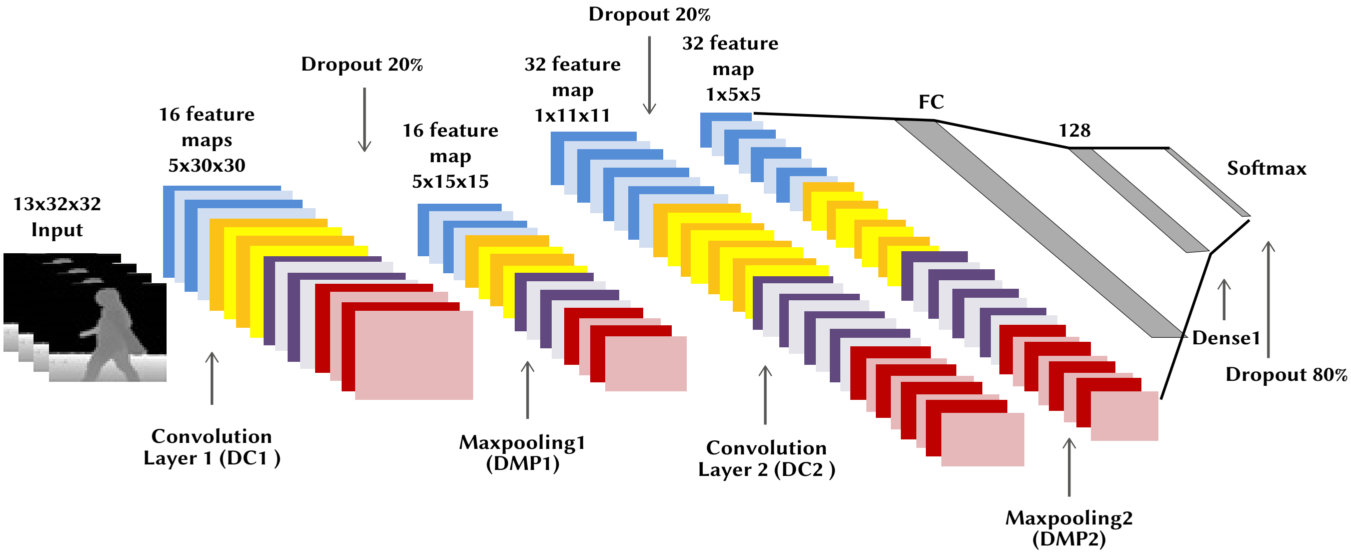


Fig. 3. The architecture of 3DCNN for spatio-temporal feature learning from depth maps.

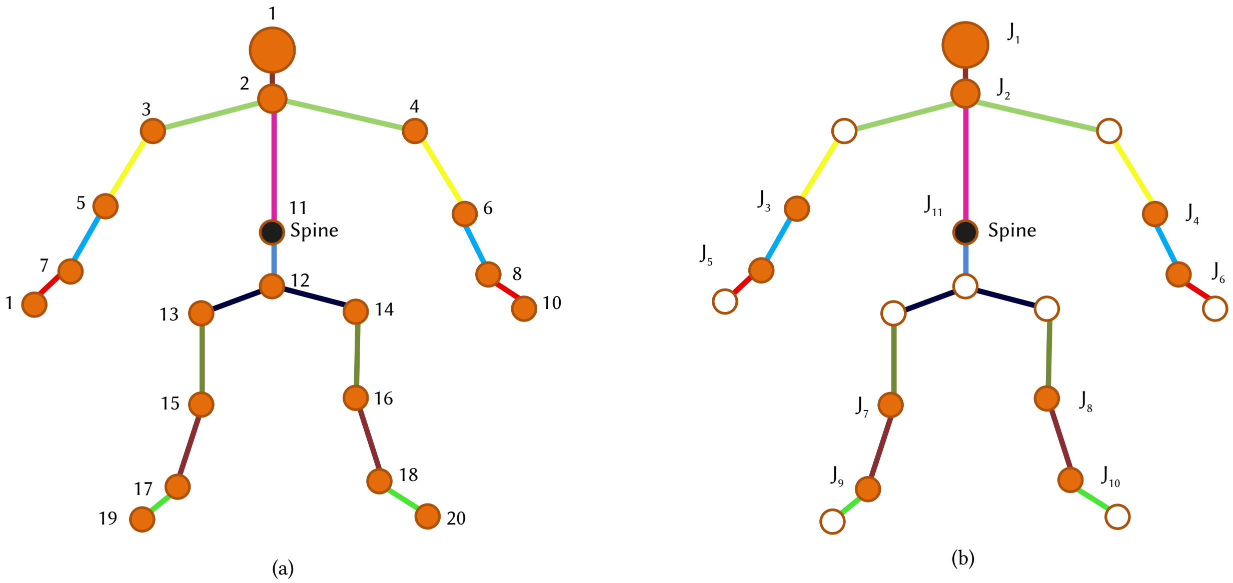


Fig. 4. 3D joints selected as per the evolutionary algorithm where J represents the positions of 11 body joints.

maxpooling layers are (2×2×2) and (1×2×2) respectively. We have used the same network settings to learn the spatio-temporal feature from the depth maps for both MSRDailyActivity3D and UTKinectAction3D datasets. The architecture of 3DCNN for spatio-temporal features learning from depth videos is given in Fig. 3.

C. Spatio-Temporal Activity Learning From Skeleton Joints Information Using LSTM

In the present section, we have extracted the set of relevant and observable features from the skeleton joint sequence and prepared a feature vector (F_T). Three features corresponding to position, spatial and temporal dimension are used to make a feature vector (F_T). In order to make F_T we used, 3D joints position (F_{3DJ}), Minkowski distance (F_{MD}) and Temporal (F_{Temp}) features. The feature vector is given in (2)

$$F_T = \{F_{3DJ}, F_{MD}, F_{Temp}\} \tag{2}$$

1. 3D Joints Positions Feature

A twenty 3D body joints corresponding to an activity captured by a kinect sensor in the dataset are shown in Fig. 4(a). However, not all joints are informative and useful for activity recognition due the unwanted noise present in the device during data capturing and the orientation of the human body. Therefore, a minimal set of joints positions are recognized using an evolutionary algorithm which determines the optimal set of 3D joints positions. For this purpose, we have used the method performed in [63] that eliminates the redundant joints positions because of closeness between the joints, and provides the optimal set of joints positions for activity recognition. For example, joints (wrist and ankles) are redundant to joints (hands and feet) and not informative for activity recognition. Using the above concept, a feature (F3DJ) is extracted having eleven 3D joints positions as shown in Fig. 4(b). The optimal set of joints position is given in (3)

$$F_{3DJ} = \{J_1, J_2, J_3\, J_4 J_5 J_6, J_7, J_8, J_9, J_{10}, J_{11}\} \tag{3}$$

Where $J_1, J_2, J_3 ...... J_{11}$ are the positions of the optimal body joints.

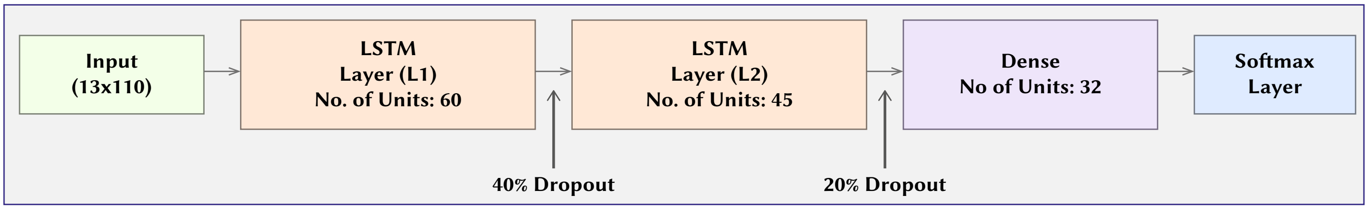


Fig. 5. Architecture of LSTM for spatio-temporal features learning from Skeletons joints Sequences.

2. Minkowski Distance Feature

To learn the spatial aspects of an activity sequence, we have utilized the concept of symmetric matrices using the 3D body joints positions. The Minkowski distance is the generalization of the both Euclidian and Manhattan distances.

Here, the Minkowski distance feature has been calculated from the optimal joints say N obtained from the equation (3). Here the value of N is 11 from the equation (3). The Minkowski distance between the two points X = (x_1, x_2, $x_{3,}$.......x_n) and Y = (y_1, y_2, $y_{3,}$.......y_n) of order r is defined in (4).

$$Dist_r(X,Y) = \left(\sum_{l=1}^{n} [x_i - y_i]^r\right)^{1/r} \tag{4}$$

Finally, a 55-dimensioal feature vector is generated corresponding to the Minkowski distance matrix.

3. Temporal Feature

To learn the temporal information from the activity sequence of 3d joints we first find the maximum and minimum values of the coordinate of any joint J_i, then for each frame in the sequence we obtain the difference between the coordinates for the frame to the maximum and minimum values of the same 3D joint for a complete sequence. For a 3D joint J_i, and 3D coordinates system (J_{ix}, J_{iy}, J_{iz}), , we obtained($J_{i,max}$, $J_{i,min}$) as given in the equation (5) and (6) respectively.

$$J_{i,max} = \frac{(\max(J_{ix_t}) - J_{ix}) + (\max(J_{iy_t}) - J_{iy}) + (\max(J_{iz_t}) - J_{iz})}{3} \tag{5}$$

$$J_{i,min} = \frac{(J_{ix} - \min(J_{ix_t})) + (J_{iy} - \min(J_{iy_t})) + (J_{iz} - \min(J_{iz_t}))}{3} \tag{6}$$

Where ⟨max (J_{iy_t}), max (J_{iy_t}), max (J_{iz_t})⟩ and ⟨min (J_{iy_t}), min (J_{iy_t}), min (J_{iz_t})⟩ are the maximum and minimum values of the coordinates of joint J_i for all the sequence respectively. Finally, a 22 dimensional temporal feature (F_{Temp}) vector is obtained by combining all the minimum and maximum values such as

$F_{TEMP} = \{J_{1,min}, J_{2,min}, ...J_{N,min}, J_{1,max}, J_{2,max}, ...J_{N,max}\}$

4. Feature Learning Using LSTM

The LSTM network is proposed by Hochreiter and Schmidhuber [62] in 1996. Not at all like other RNN, LSTM protects and keeps up the errors so that they can be effortlessly backpropagated through layers, which also makes LSTMs valuable for time series predictions and makes the model to continuously learn using a wide number of time steps. A LSTM network contains input gate, output gate, forget gate and memory units in the recurrent layer. The memory units contain the memory cells which are used to maintain the temporal state of the network using self-connections. These gates performed the function according to the received signal and pass or block the information based on its strength. The gate weights are used to filter these signals. These weights act in the same way as ordinary NN input and hidden states as they are learned all through the RNN learning prepare.

The input gate is utilized to assure that the all of data included to the cell stated is vital and not redundant and it controls the flow of input into the memory cells. The careful selection of valuable data from the current cell state and showing it as an output is done with the output gate. A forget gate is outlined for annihilating information from the current cell state. The LSTM execution is optimized by the evacuation of any meaningless information that is not required by the LSTM to get it things or to evacuate any information that is not important anymore. An LSTM network performs a mapping from the input sequence X = {x_1, x_2, x_3..... x_t} to the output sequence Y = {y_1, y_2, y_3..... y_t} using the network activations reclusively from t = 1 to T.

The space time feature vector (F_T) given in (2) is used to train the LSTM network, which contains two LSTM layers with 60 neurons and 45 neurons respectively. Two dropout layers are used after each LSTM layer with 20 % dropout in each case. Then a dense layer is used with 32 neurons. Finally a softmax layer is used for output classifications. The rectified linear unit ‘ReLU’ activation function is used in each LSTM layer. The architecture of LSTM is given in Fig. 5.

D. Score Prediction Using SVM

The SVM classifier is based on the kernels [64]. The main objective of this classifier is to map the input samples into the higher dimensional feature space, and insert a hyperplane that distinctly classify the input samples. SVM performs both types of classification, linearly and non-linearly, with the help of different kernels. Some of the kernels of SVM are linear, radial bases function (RBF), polynomial, etc. In our work, one-vs-all SVM classifier with radial basis function (rbf) kernel is trained using the training features extracted from the second last layer (layer before softmax) of the network for each models independently. To make one-vs-all SVM classifier, the value of decision_function_shape parameter has been set to ‘ovr’. During the testing, the corresponding test features have been extracted for each test sequence and applied to the trained SVM model. Then, the classification probability scores of each test sequence are calculated. Finally, the obtained probability scores for each test sequence of the three models are fused together for optimization using evolutionary algorithm.

E. Score Optimization Using Evolutionary Algorithm

In this section, we used two optimization algorithms, GA and PSO separately to find the optimized class score of each test activity.

1. Score Optimization Using Genetic Algorithm

Genetic Algorithm is developed by Goldberg [65] in 1989. GA is a heuristic search method to solve both unconstrained and constrained optimization problem which is based on the process of natural selection. GA iteratively changes the population of individual solutions. The algorithm picks individuals from the current population in each step and utilizes them as a parent to create the population for the next generation. The population generates the optimal solution after the continuous iterations. The process of the genetic algorithm starts with an initial population and the choice of the initial population generally depends on the optimization problem. The next process determines the selection of the proportion of existing population to evaluate the objective function. The ratio of the current population is taken to breed the new generation and fitness function is used to select the individual solution. Then the crossover and mutation process is used to generate a second level population of solutions. In

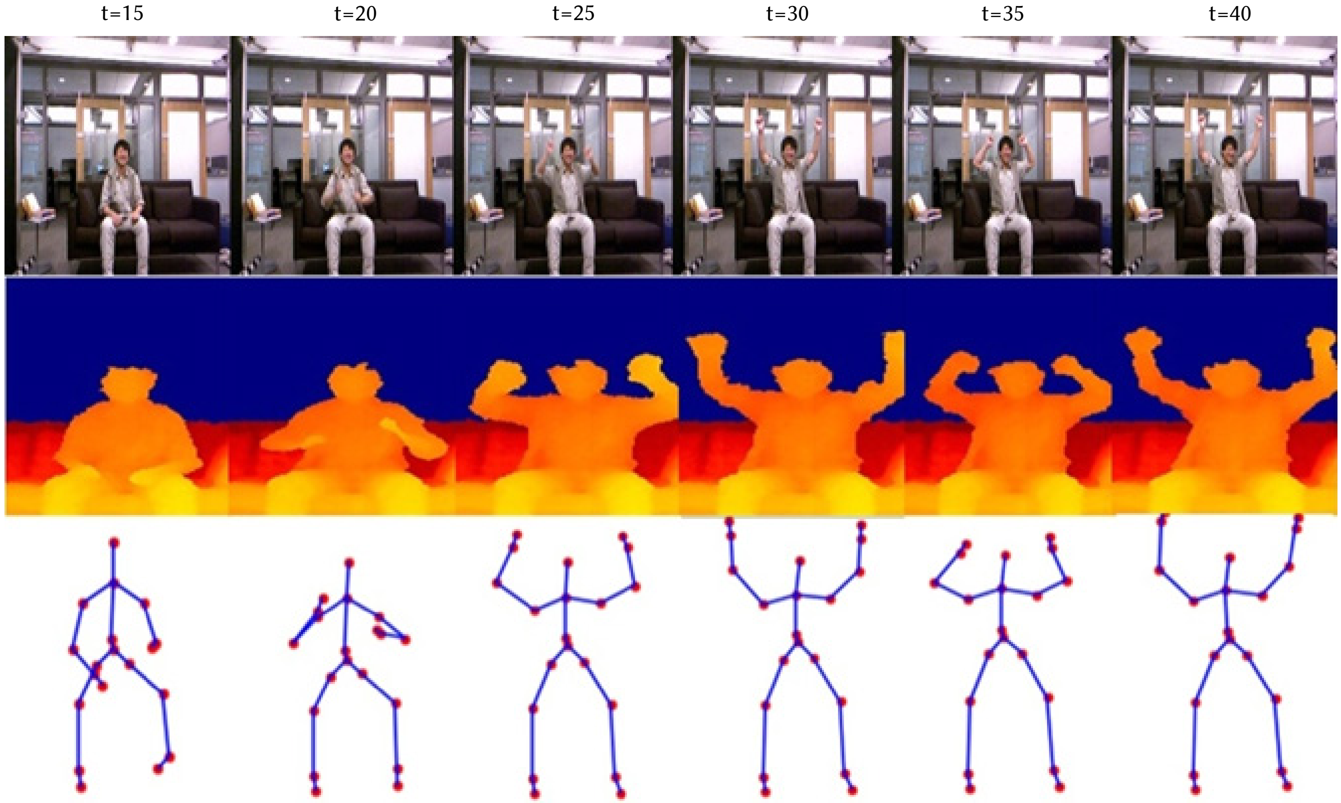


Fig. 6. Sample frames of 'cheer up' in MSRDailyActivity3D dataset at different time stamp in the three formats RGB, Depth and Skeleton.

the crossover process the selected parents are combined to generate the new population. The mutation process is used to produce the children by making the random change in parents. Similarly, the same process start again by inserting the new generation into the selection process and the algorithm repeats itself until some termination condition is satisfied. Some of the termination conditions are when the objective function reaches some threshold condition, fixed number of generations reached, or time limitation etc.

In this work, GA is used to optimize the probability scores for each activity sequences obtained from the trained SVM classifier for each modality. The optimal unbounded weights have been recorded against the highest accuracy. In addition to GA optimization, we have also used the PSO algorithm to optimize the probability score and compared the results of both optimizations.

2. Score Optimization Using PSO

The PSO is a well-known optimization algorithm based on population developed [66]. The PSO attempts to find the optimal solution of the problem using the population of particles. The basic principle of PSO is one in which each particle in the swarm represents an individual. And the combination of particles is a swarm. The PSO begins working in parallel, with a collection of particles, and reaches to the optimal solution using the current velocity, its prior best velocity and velocity of its neighbor particles. Some of the key features of the PSO are simple, effective, easy to implement and it does not required gradient information. The search space is considered as the solution space in PSO and a position in the search space signifies the solution of the problem. Different particles move in the search space with its velocity to find the optimal solution in the search space. The particle movement at each iteration is described in equation (7) and (8)

$$X_p(t+1) = X_p(t) + v_p(t) \tag{7}$$

$$v_p(t+1) = \omega v_p(t) + a_1 r_1 \left(xbest_p(t) - X_p(t)\right) + a_2 r_2 \left(ybest_p(t) - X_p(t)\right) \tag{8}$$

Where $X_p(t)$ is the position of the particle p at time t, $v_p(t)$ is the velocity of the particle at time t, $xbest_p(t)$ is the best position of the particle found by itself. $ybest_p(t)$ is the best position determined by all the swarm. ω is the inertia weight, a_1 a_2 are the two acceleration coefficients and r_1 r_2 are the two random variables whose value ranges between 0 and 1.

In this phase, the obtained probability scores are also optimized using PSO. The optimal weights corresponding to highest accuracy have been achieved after the successive iteration of the PSO algorithm.

IV. Experimental Results and Discussions

In this paper, we used the MSRDailyActivity3D and UTKinectAction3D datasets to train and test our proposed approach. Both datasets were captured using Kinect Senor. Activities in both datasets are synchronized in all formats such as RGB, Depth, and Skeleton. To train our model we used Intel Core i7 8th generation, 2.6 GHz processor with 16 GB of RAM, on Ubuntu 16.04 LTS (Linux) operating system having 2 GB 940MX NVIDIA GPU support. To implement CNN and RNN networks, we use the Keras Deep learning framework with version 2.2.4, and for implementing the Genetic algorithm optimization, we used MATLAB version 18a.

A. MSRDailyActivity3D Dataset

The MSRDailyActivity3D dataset [67] was captured with the help of Kinect sensor. It contains 16 human activities such as: 'drink', 'eat', 'read book', 'call cellphone', 'write on a paper', 'use laptop', 'use vacuum cleaner', 'cheer up', 'sit still', 'toss paper', 'play game', 'lie down on sofa', 'walk', 'play guitar', 'stand up', 'sit down'. The dataset is captured by ten subjects, out of them five are male and five are

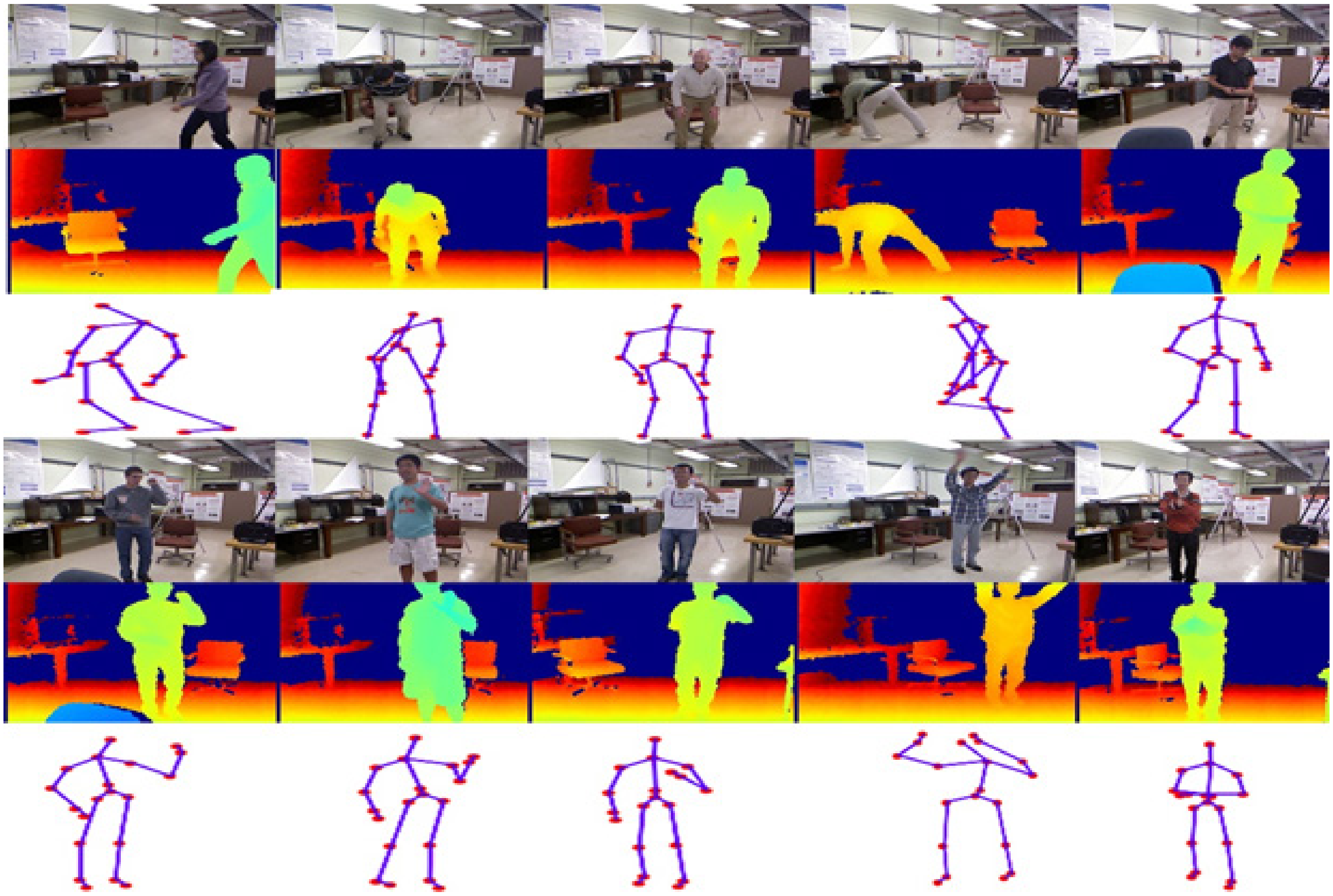


Fig. 7. Some sample frames of different activities of the UTKinectAction3D dataset in three different formats, RGB, Skeleton and Depth.

females. Each activity is captured inside the living room and there is a sofa in the scene, this means there is an interaction of subject to the object during activity capturing. Each activity is carried out by each subject once in sitting position and once in standing position. Each activity is captured in three different forms, RGB, Depth map and Skeleton joints. The file format of RGB, Depth maps and Skeleton joints are avi, bin and txt files. There are 16(activity) × 10 (person) × 2 (position) = 320 files for each modality. Therefore, 320 × 3 = 960 files in total. Fig. 6 shows a 'cheer up' activity of MSRDailyActivity3D at different time stamps in three different formats RGB, Depth and Skeleton dataset samples.

B. UTKinectAction3D Dataset

The UTKinectAction3D dataset was collected by the Microsoft Foundation Research in 2012. It contains 10 indoor actions performed by ten subjects, out of them 9 are males and 1 is female. The captured actions are 'carry', 'walk', 'sitdown', 'standup', 'push', 'pull', 'throw', 'pickup', 'wavehands', and 'claphands'. Each subject performs each activity twice. The data is captured in three formats, RGB videos, the depth map, and skeleton joints, and all activities are synchronized in all formats. The second carry activity sequence is not given hence there are total 199 activity sequences that exist in each modality. For the experimental point of view, a total of 200 sequences are used in this work. There are 10(action) × 10(subjects) × 2(frequency) = 200 activities in each case. In total 200 × 3 = 600 activities corresponding to RGB, depth, and skeleton data. The file format of the RGB video frames, depth map, and skeleton joint positions are jpg, bin, and txt file. The dataset is captured using a stationary Kinect camera. Fig. 7 shows the ten actions of UTKinectAction3D dataset.

C. Experimental Results on MSRDailyActivity3D Dataset

During the implementation phase, we extract the features from the RGB video sequences using a 3D Convolutional Neural Network, from the depth maps using second 3D Convolutional Neural Network and from the skeleton joints by a LSTM network for later classification by SVM. Different input cube size such as (13×50×50), (14×50×50), (15×50×50), (16×50×50) are applied to first 3DCNN and obtained the best spatio-temporal features at input size (16×50×50). Similarly for depth data, the input cubes size (13 × 32 × 32), (14 × 32 × 32), (15 × 32 × 32) and (16 × 32 × 32) have been tried to the second 3DCNN and best features are extracted at (13 × 32 × 32). To train both 3DCNN, we used a 6×10^{-4} learning rate by a decay factor that decreases with increases in the number of epochs. An Adam optimizer is used with a categorical cross-entropy loss function.

For validation of our proposed approach, we used Leave-One-User-Out cross-validation (LOUOCV). The dataset is divided into two subsets, one for the training set and another testing set. Ten folds cross-validation is used, and in each fold, nine users are used for training purpose and one user for testing. In (LOUO) cross-validation technique 9 × 16 × 2 = 288 activity sequences are used in training, and 1 × 16 × 2 = 32 activity sequences are used to perform the test in each fold. The class probability scores of each activity sequence are recorded during the testing phase.

In GA optimization, the class probability scores obtained from all three models during the cross-validation step are fused together. Here, we used the roulette wheel procedure for chromosomes selection. For the optimization, the mutation probability is set to 0.01, and the crossover rate is set to 0.8. After the successive iterations, the GA optimizes the weights and increases the classification accuracy. The GA optimization is performed with varying population sizes from 10

TABLE I. Optimized GA Results for MSRDailyActivity3D Dataset

No. of steps	No of Iteration per step	Population Size	Accuracy	w_1	w_2	w_3
1	72	10	83.44	5.429	7.367	0.854
2	51	20	84.69	9.8	9.473	6.793
3	**87**	**30**	**85.94**	**4.819**	**0.277**	**7.854**
4	76	40	83.44	2.697	3.82	4.798
5	82	50	84.37	3.339	5.197	3.763
6	68	60	84.10	2.982	-20.614	36.706
7	57	70	83.12	1.565	3.714	6.445
8	76	80	84.10	1.211	-3.296	8.612
9	67	90	83.43	1.211	-3.296	8.612
10	74	100	83.13	4.814	3.471	6.662

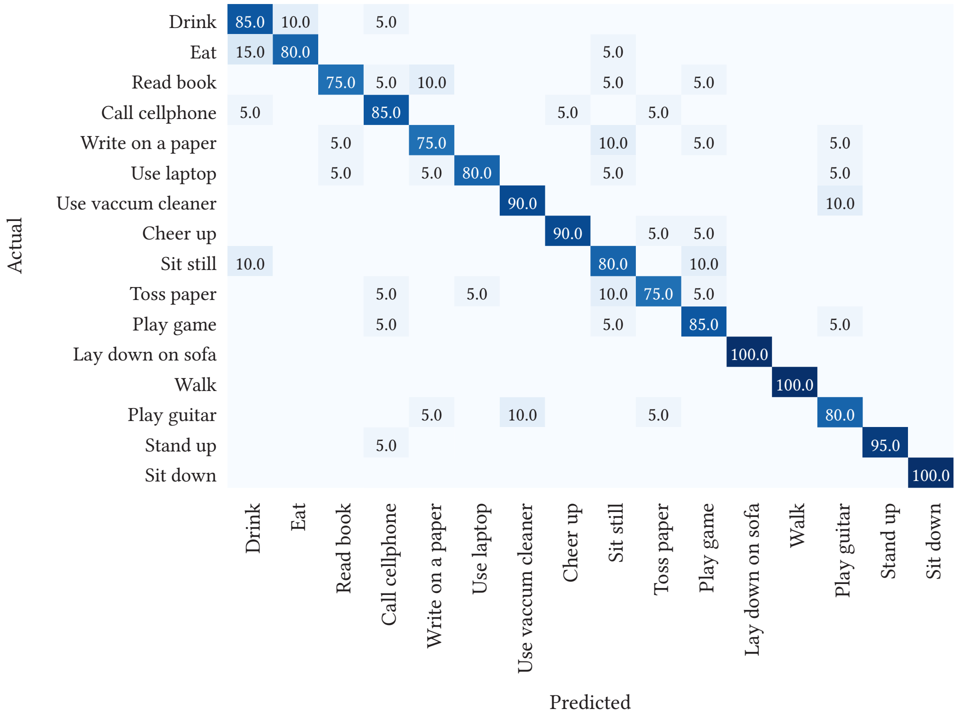


Fig. 8. Confusion Matrix of our proposed approach for MSRDailyActivity3D Dataset.

to 100. The maximum classification accuracy of **85.94%** is achieved when population size is 30 at step 3 from Table I, and corresponding optimum weight values have been recorded as w_1 = 4.819, w_2 = 0.277 and w_3 = 7.854.

It can also be observed from Table I, that performance decreases with increase in the population size, and the maximum classification accuracy is achieved at population size 30. The values of learned weight, classification accuracy, number of iterations per step along with population size, are given in Table I.

The confusion matrix of our proposed approach for the MSRDailyActivity3D dataset is given in Fig. 8. Interpretations of confusion matrix clearly show that the recognition accuracy of activities 'read book', 'write on a paper' and 'toss paper' are comparatively low due to similarity in these three activities. Likewise, the activities 'drink' and 'eat' are identical to each other due to the similar body part movement, therefore containing more confusion among them, and our model is predicting 15 percent of the times 'eat' activity as 'drink' activity. The recognition accuracy of the activities belonging to the second half of the confusion matrix shown in Fig. 8 are better than the activities belonging to the first half of the confusion matrix, since the confusion between the activities in the first half is more as compare to the activities belonging to the second half. For example, the activities 'sit down', 'walk', 'lay down on sofa' are recognized 100%, 'stand up' is also approximate to 100% except for one sample which is being misclassified as 'call cell phones', while the activities 'drink', 'eat', 'read book', 'call cell phones', 'write on a paper', 'use laptop' have lower recognition accuracies. One of the reasons for this is that, the activities such as 'read book', 'write on a paper', 'toss paper' are happening with the contact of the external object. Therefore, their recognition accuracies are less as compared to the activities in which no external object is being used.

After the above discussion, it is clear that all the activities have good recognition accuracies, except those that have higher confusion among them and those that are happening with the contact of external objects, which proves that our proposed approach learns the structural and motion information as well as relevant features from the input sequences in a better way.

In addition to applying the GA, we have also used PSO algorithm as an alternative method to optimize class scores. We repeat the optimization process ten times with unbounded weights. During the optimization process, PSO optimized the weights and improves the classification accuracy. The optimal weights and their corresponding classification accuracy are given in Table II.

TABLE II. Optimized PSO Result on MSRDailyActivity3D Dataset

No. of steps	Iteration per step	Classification Accuracy	w_1	w_2	w_3
1	31	81.56	.280	.3167	1.3071
2	30	82.19	.4702	-.9635	2.3384
3	24	83.13	.3974	.4310	1.8351
4	28	82.50	.1092	-1.5660	1.8581
5	33	82.81	2.0110	-1.1847	7.0962
6	29	82.81	0.3044	.3379	1.4145
7	27	81.87	0.3412	1.3601	2.3577
8	**30**	**83.75**	**1.0431**	**-0.2076**	**3.2864**
9	32	82.50	2.7499	1.6599	9.7201
10	24	82.19	1.8581	-.3211	5.8407

TABLE III. Performance Comparison of the Proposed Approach with Other Methods on MSRDailyActivity3D Dataset

Dataset Modality	Methods	Accuracy (%)
Depth	Only LOP feature [67]	42.5
	Histogram of oriented 4Dnormals [68]	80.0
	Depth Cuboid similarity feature (DCSF)[69]	83.6
Skeleton	NBNN [70]	53
	NBNN + time [70]	60
	NBNN + parts [70]	60
	Only Joint Position feature[67]	68
	NBNN + parts + time [70]	70
	Distance + Temporal features [71]	73.43
	mean 3D joints [72]	73.75
	SVM + FTP feature [67]	78
Hybrid	MHI+SAE feature [58]	74.6
	Actionlet Ensemble [67]	85.75
	Our Proposed Approach with GA	**85.93**
	Our Proposed Approach with PSO	**83.75**
	DCSF + joint [69]	88.2

TABLE IV. Optimized GA Results on UTKinectAction3D Dataset

No. of steps	No of Iteration per step	Population Size	Accuracy	w_1	w_2	w_3
1	47	10	95.5	7.652	14.132	10128
2	51	20	96	1.367	4.121	3.835
3	48	30	95.5	4.204	8.187	5.61
4	49	40	95.5	2.308	8.063	9.267
5	**54**	**50**	**96.5**	**1.632**	**6.529**	**8.008**
6	42	60	96	4.352	7.416	5.262
7	44	70	95.5	2.344	6.142	4.671
8	39	80	95.5	2.277	9.889	11.913
9	41	90	95	1.054	1.711	1.79
10	36	100	95.5	1.464	7.134	8.46

It can be concluded from Table II that, highest classification accuracy has been achieved in step 8. The PSO optimized weights are w_1 = 1.0431, w_2 = 0.2076 and w_3 = 3.2864, the corresponding classification accuracy is **83.75%**.

Table III contains the state-of-art methods along with their accuracy on the MSRDailyActivity3D benchmark dataset. The comparison of the recognition accuracy of our proposed method against the mentioned state-of-art methods are given in Table III and it can also be seen from Table III that the proposed approach with GA optimization has comparable accuracy with the mentioned state-of-art results.

D. Experimental Results on UTKinectAction3D Dataset

To confirm the efficacy of our method, we have evaluated our proposed approach on another benchmark UTKinectAction3D dataset. For the experiment point of view, we used the same experimental setup as used in the MSRDailyActivity3D dataset. For validation, we use leave one user out (LOUO) cross-validation in ten rounds, in which one user is removed from training and used as testing. The process is continuing for all users. In LOUO cross-validation method, 10 × 9 × 2 = 180 activities sequences are used in training, and 10 × 1 × 2 = 20 activities sequence are used in testing to test each user in each round. The probability scores of each activity sequence are recorded and fused together. Next, the GA optimizes the weights and improves the classification accuracy during successive iterations. It can be observed from Table IV, that the maximum classification accuracy 96.50% is achieved at population size=50 in step 5 from Table III, and the corresponding optimum weights are w_1 = 1.632, w_2 = 6.529 and w_3 = 8.008.

TABLE V. Optimized PSO Result on the UTKincecAction3D Dataset

No. of steps	Iteration per step	Classification Accuracy	w_1	w_2	w_3
1	42	94.5	2.3633	6.4472	-0.2244
2	**39**	**95**	**1.3515**	**3.7570**	**-1.0144**
3	37	94.5	1.4766	5.2991	0.3548
4	41	94.5	2.5806	5.5602	2.4345
5	35	94	1.5175	5.4437	0.6930
6	37	93.5	0.9930	3.4147	-0.6840
7	43	93.5	4.0628	5.8333	1.8558
8	38	94	1.6915	3.1000	-0.2983
9	41	93.5	0.3260	1.0256	0.4786
10	39	94.5	1.3927	1.8255	0.1039

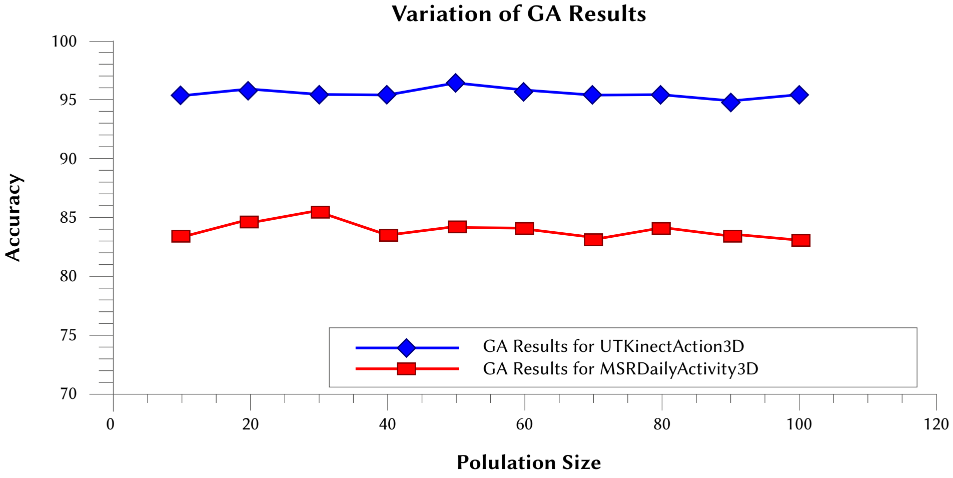
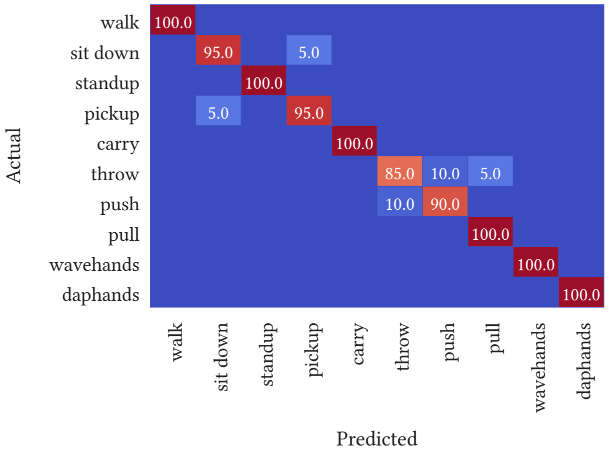


Fig. 10. Variation of GA accuracies on UTKinectAction3D and MSRDailyActivity3D datasets at different population size.

The confusion matrix of the UTKinectAction3D dataset is given in Fig. 9. It can be conclude from Fig. 9 that the activity ‘throw’ has comparatively less recognition accuracy due to the high degree of confusion with push activity. It can also be illustrated from the Fig. 9 that confusion occurs between the activities ‘sitdown’ and ‘pickup’ while most of the activities are 100% classified. Therefore, it proves that our proposed approach is sufficient to learn spatial and motion features from the activity sequences. The performance with the UTKinectAction3D dataset has been compared against state-of-art-methods.



	walk	sit down	standup	pickup	carry	throw	push	pull	wavehands	daphands
walk	100.0									
sit down		95.0		5.0						
standup			100.0							
pickup		5.0		95.0						
carry					100.0					
throw						85.0	10.0	5.0		
push						10.0	90.0			
pull								100.0		
wavehands									100.0	
daphands										100.0

Fig. 9. Confusion matrix of UTKinectAction3D dataset.

Similarly, we have also applied the PSO algorithm on the UTKinectAction3D dataset to optimize the class scores of each test activity. The PSO optimization is performed in ten steps. The obtained results are given in the Table V.

Based on Table V, the highest classification accuracy is achieved in step 2. The PSO optimized weights are $w_1 = 1.3515$, $w_2 = 3.7570$ and $w_3 = -1.0144$ and the corresponding classification accuracy is **95.0%**.

Fig. 10 describes how the GA accuracy varies at different population size for both UTKinectAction3D and MSRDailyActivity3D datasets. The performance of UTKinectAction3D dataset has been compared against state-of-art-methods. Table VI displays the comparison results. Our proposed approach with GA optimization gives the highest classification accuracy over all other approaches.

TABLE VI. Performance of Our Proposed Approach on UTKinectAction3D Dataset, Compared to the State-of-art-approaches

Methods	Accuracy (%)
Xia. et al., (2012) [5]	90.92
Wang. et al., (2015) [49]	90.91
Liu et al., (2015) [73]	92.00
Liu. et al., (2016) [74]	96.00
Proposed Approach with GA	**96.50**
Proposed Approach with PSO	**95**

E. Error Analysis

The above discussion indicates that, during the space-time features learning using 3DCNN, we have taken the fixed number of frames in the input cube because the length of the frames in every video varies from 51 to 553 for MSRDailyActivity3D dataset and 5 to 120 for UTKinectAction3D dataset respectively, therefore some loss of information occurs.

In MSRDailyActivity3D dataset, the total false positive rate (error) is 0.141. This is because the upper half of the activities in the confusion matrix shown in Fig. 6 such as 'drink', 'eat', 'read book', 'write on a paper', 'use laptop', 'toss paper', 'sit still' and 'play guitar' have FPR 0.3125 while the remaining below half of the activities have FPR only 0.06875. The reason is that the activities in the first half of the confusion matrix are similar to each other and have more confusion compared to the activities in the second half. For example, the activities 'read book', 'write on a paper', and 'use laptop' have similar body movement with an external object. Also, the activities 'drink' and 'eat' are more confusing due to higher homogeneity levels. Furthermore, the false-positive rate (FPR) of our proposed approach for the UTKinectAction3D dataset is 0.035, out of which 0.025 error is mainly due to the two activities 'push' and 'throw'. The reason behind this is that, the uniformity between the activities 'push' and 'throw' and another reason is that the network learns almost similar motion information for both activities.

V. Conclusion and Future Work

In this paper, a Deep Multi-Model approach has been proposed for activity recognition, which mainly consists of three steps. Firstly, we extract spatial and temporal features using two 3DCNN and a LSTM networks from RGB, Depth, and skeleton information respectively. Secondly, three SVM classifiers are used to generate the class probability scores of each test activity in all formats. Finally, the class scores obtained from each modality are fused and optimized by a genetic algorithm for activity recognition. The proposed approach automatically learns high-level features from input data for each modality using spatial-temporal convolutional neural networks, and a LSTM network. Our proposed approach also uses RGB, Depth and skeleton information and gives better performance than using each modality separately. An optimization-based score fusion technique is presented to take full advantage of class label decisions from different aspects. For this purpose, we use the genetic algorithm (GA). We have evaluated our proposed approach on two benchmarks, MSRDailyActivity3D and, UTKinectAction3D datasets, and have achieved 85.94% and 96.50% recognition accuracies, respectively. The obtained results are comparable over the state-or-art-methods. Moreover, in our proposed approach, the three channels can run in parallel, therefore, it can also run on the multi-core CPU systems to save time. Hence this arrangement makes our method more efficient and fast.

Our proposed method is an application of the human-machine interface. It can be used to recognize the normal activities, abnormal activity and patient security and monitoring inside the living room etc. Despite the several advantages, our proposed system has two drawbacks. Firstly, the depth camera can capture the frames only from 4 to 11 feet. Secondly, the depth video sequences must be captured completely prior to test over the network. In future work, these challenges may be resolved.

Acknowledgment

We are grateful to the College of Engineering Roorkee, India, and UTU Dehradun, India, for providing excellent research facility to carry out this research work.

References

[1] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," in IEEE transactions on pattern analysis and machine intelligence, vol.40, no. 3, pp.667-681, 2017.

[2] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," in IEEE Transactions on Image Processing, vol. 27, no. 6, pp.2842-2855, 2018.

[3] J. K. Aggarwal, and M. S. Ryoo, "Human activity analysis: A review," in ACM Computing Surveys (CSUR), vol. 43, no. 3, pp.1-43, 2011.

[4] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 2010, pp. 9-14.

[5] L. Xia, C.C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 2012, pp. 20-27.

[6] X. Yang,, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in Proceedings of the 20th ACM international conference on Multimedia, Nara, Japan, 2012, pp. 1057-1060.

[7] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 2013, pp. 486-491.

[8] A. Chaaraoui, J. Padilla-Lopez, and F. Flórez-Revuelta, "Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices," in Proceedings of the IEEE international conference on computer vision workshops, Sydney, NSW, Australia, 2013, pp. 91-97.

[9] S. Siddiqui, M. A. Khan, K. Bashir, M. Sharif, F. Azam, and M. Y. Javed, "Human action recognition: a construction of codebook by discriminative features selection approach," in International Journal of Applied Pattern Recognition, vol. 5, no. 3, pp.206-228, 2018.

[10] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and RGB data," in Pattern Recognition Letters, vol. 131, pp. 293-299, 2020.

[11] K. Khoshelham, and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," in Sensors, vol. 12, no. 2, pp. 1437-1454, 2012.

[12] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers and S. A. Velastin, "Exploiting deep residual networks for human action recognition from skeletal data," in Comput Vis Image Underst, vol. 170, pp. 51–66, 2018.

[13] S. Yang, J. Yang, F. Li, G. Fan and D. Li, "Human Action Recognition Based on Fusion Features," in International Conference on Cyber Security Intelligence and Analytics, 2019, pp. 569–579.

[14] A. Jalal, M. Z. Uddin, and T. S. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," in IEEE Transactions on Consumer Electronics, vol. 58, no. 3, pp.863-871, 2012.

[15] M. Khan, T. Akram, M. Sharif, N. Muhammad, M. Javed and S. Naqvi, "An improved strategy for human action recognition; experiencing a cascaded design," in IET Image Processing, vol 14, no. 5, pp. 818-829, 2019.

[16] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", in IEEE conference on computer vision and pattern recognition, Los Vegas, NV, USA, 2016, pp. 770–778.

[17] L. Bi, D. Feng and J. Kim, "Dual-path adversarial learning for fully convolutional network (FCN)-based medical image segmentation," in Visual Computers, vol. 34, no. 6, pp. 1–10, 2018.

[18] M. Rashid, M. A. Khan, M. Sharif, M. Raza, M. M. Sarfraz and F. Afza, "Object detection and classification: a joint selection and fusion strategy of deep convolutional neural network and SIFT point features," in Multimedia Tools and Applications, vol. 78, no. 12, pp.15751–15777, 2019.

[19] F. Zhou, Y. Hu and X. Shen, "Msanet: multimodal self-augmentation and adversarial network for RGB-D object recognition," The Visual Computers, vol. 35, no. 11, pp. 1583-1594, 2019, https://doi.org/10.1007/s00371-018-1559-x

[20] I. Gogić,, M. Manhart, I. S. Pandžić and J. Ahlberg,, "Fast facial expression recognition using local binary features and shallow neural networks," in The Visual Computer, vol. 36, no. 01, pp.1-16, 2018.

[21] M. Sharif, M. A. Khan, M. Faisal, M. Yasmin and S. L. Fernandes, "A framework for offline signature verification system: Best features selection approach," in Pattern Recognition Letters, 2018.

[22] K. K. Verma, B. M. Singh and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system", in International Journal of Information Technology, 2019, pp. 1-14.

[23] G. I. Parisi, "Human Action Recognition and Assessment via Deep Neural Network Self-Organization," in Modelling Human Motion, pp. 187-211, 2020.

[24] X. X. Niu and C. Y. Suen, "A novel hybrid CNN–SVM classifier for recognizing handwritten digits," in Pattern Recognition, vol. 45, no. 4, pp. 1318-1325, 2012.

[25] D. X. Xue, R. Zhang, H. Feng and Y. L. Wang, "CNN-SVM for microvascular morphological type recognition with data augmentation," in Journal of medical and biological engineering, vol. 36, no. 6, pp. 755-764, 2016.

[26] A. B. Sargano, X. Wang, P. Angelov and Z. Habib, "Human action recognition using transfer learning with deep representations," in 2017 International joint conference on neural networks (IJCNN), Anchorage, AK, USA, 2017, pp. 463-469.

[27] T. Jiang, Z. Zhang and Y. Yang, "Modeling coverage with semantic embedding for image caption generation," in The Visual Computers, vol. 35, no. 11, pp. 1655-1665, https: //doi.org/10.1007/s00371-018-1565-z

[28] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei and S. Savarese, "Social lstm: human trajectory prediction in crowded spaces," in IEEE conference on computer vision and pattern recognition, 2016, pp 961–971

[29] I. Sutskever, O. Vinyals Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104–3112.

[30] K.K. Verma, B. M. Singh, "Deep Learning Approach to Recognize COVID-19, SARS and Streptococcus Disease from Chest X-Ray Images," in Journal of Scientific and Industrial Research, vol. 80, no. 01, pp. 51-59, 2021.

[31] J. Cong and B. Zhang, "Multi-model feature fusion for human action recognition towards sport sceneries," in Signal Processing: Image Communication, 2020.

[32] E. Zhou and H. Zhang, "Human action recognition towards massive-scale sport sceneries based on deep multi-model feature fusion," Signal Processing: Image Communication, vol. 84, 2020.

[33] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[34] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: a large video database for human motion recognition," in 2011 International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 2556-2563.

[35] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 2015, pp. 961-970.

[36] A. B. Sargano, P. Angelov and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," in Applied sciences, vol. 7, no. 01, 2017.

[37] H. Wang, and C. Schmid, "Action recognition with improved trajectories," in Proceedings of the IEEE international conference on computer vision, Sydney, NSW, Australia, 2013, pp. 3551-3558.

[38] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, vol. 25, pp. 1097-1105, 2012.

[39] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568-576.

[40] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1725-1732.

[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 2015, pp. 4489-4497.

[42] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 2016, pp. 1933-1941.

[43] G. Varol, I. Laptev and C. Schmid, "Long-term temporal convolutions for action recognition," in IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 6, pp.1510-1517, 2017.

[44] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," in IEEE Access, vol. 6, pp.1155-1166, 2017.

[45] K. K. Verma, B. M. Singh, H. L. Mandoria and P. Chauhan, "Two-Stage Human Activity Recognition Using 2D-ConvNet," in International Journal of Interactive Multimedia & Artificial Intelligence, vol. 6, no 2, pp. 135-135, 2020.

[46] Z. Li, Z. Zheng, F. Lin, H. Leung and Q. Li, "Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN," in Multimedia Tools and Applications, vol. 78, no. 14, pp.19587-19601, 2019.

[47] P. Wang, W. Li, Z. Gao, C. Tang and P. O. Ogunbona, "Depth pooling based large-scale 3-d action recognition with convolutional neural networks," IEEE Transactions on Multimedia, vol. 20, no. 5, pp.1051-1061, 2018.

[48] C. Chen, R. Jafari and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2015, pp. 1092-1099.

[49] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang and P. Ogunbona, "Deep convolutional neural networks for action recognition using depth map sequences," arXiv preprint arXiv:1501.04686, 2015.

[50] V. Megavannan, B. Agarwal and R. V. Babu, "Human action recognition using depth maps," in 2012 International Conference on Signal Processing and Communications(SPCOM), Bangalore, India, 2012, pp. 1-5.

[51] Y. Han, S. L.Chung, Q. Xiao, W. Y. Lin and S. F. Su, "Global Spatio-Temporal Attention for Action Recognition based on 3D Human Skeleton Data," in IEEE Access, vol. 8, pp. 88604-88616, 2020.

[52] B. Ren, M. Liu, R. Ding and H. Liu, "A Survey on 3D Skeleton-Based Action Recognition Using Learning Method," arXiv preprint arXiv:2002.05907, 2020.

[53] R. Saini, P. Kumar, P. P. Roy and D. P. Dogra, "A novel framework of continuous human-activity recognition using kinect," in Neurocomputing, vol. 311, pp.99-111, 2018.

[54] B. Chikhaoui, B. Ye and A. Mihailidis, "Feature-level combination of skeleton joints and body parts for accurate aggressive and agitated behavior recognition," in Journal of Ambient Intelligence and Humanized Computing, vol. 8, no. 6, pp.957-976, 2017.

[55] A. Shahroudy, J. Liu, T.T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 2016, pp. 1010-1019.

[56] Y.Gu, X Ye, W. Sheng, Y. Ou and Y. Li, "Multiple stream deep learning model for human action recognition," in Image and Vision Computing, vol. 93, 2020.

[57] P. Khaire, P. Kumar and J. Imran, "Combining CNN streams of RGB-D and skeletal data for human activity recognition," in Pattern Recognition Letters, vol. 115, pp.107-116, 2018.

[58] A. Tomas and K. K. Biswas, "Human activity recognition using combined deep architectures," in 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), Singapore, 2017, pp. 41-45.

[59] E. P. Ijjina and K. M. Chalavadi, "Human action recognition in RGB-D videos using motion sequence information and deep learning," in Pattern Recognition, vol. 72, pp. 504-516, 2017.

[60] C Zhao, M. Chen, J. Zhao, Q. Wang and Y. Shen, "3D Behavior Recognition Based on Multi-Modal Deep Space-Time Learning," in Applied Sciences, vol. 9, no. 4, pp.716, 2019.

[61] S. Ji, W. Xu, M. Yang and K. Yu, "3D convolutional neural networks for human action recognition," in IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 01, pp. 221-231, 2012.

[62] S. Hochreiter and J. Schmidhuber, "Bridging long time lags by weight guessing and "Long Short-Term Memory"," in Spatiotemporal models in biological and artificial systems, vol. 37, pp. 65-72, 1996.

[63] S. Gaglio, G. L. Re and M. Morana, "Human activity recognition process using 3-D posture data," in IEEE Transactions on Human-Machine Systems, vol. 45, no. 05, pp.586-597, 2014.

[64] O. Chapelle, V. Vapnik, O. Bousquet and S. Mukherjee S, "Choosing

multiple parameters for support vector machines," in Machine learning, vol. 46, no. 1-3, pp. 131-59, 2002.

[65] D. E. Goldberg, B. Korb and K. Deb, "Messy genetic algorithms: Motivation, analysis, and first results," in Complex systems, vol. 3, no. 05, pp.493-530, 1989.

[66] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proceedings of ICNN'95-International Conference on Neural Networks, Perth, Australia, 1995, pp. 1942-1948.

[67] J. Wang, Z. Liu, Y.Wu and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 1290-1297.

[68] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in Proceedings of the IEEE conference on computer vision and pattern recognition, Portland, OR, USA, 2013, pp. 716-723.

[69] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in Proceedings of the IEEE conference on computer vision and pattern recognition, Portland, OR, USA, 2013, pp. 2834-2841.

[70] L. Seidenari, V. Varano, S. Berretti, A. Bimbo and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland OR, USA, 2013, pp. 479-485.

[71] Y. Hbali, S. Hbali, L. Ballihi and M. Sadgal, "Skeleton-based human activity recognition for elderly monitoring systems," in IET Computer Vision, vol. 12, no. 01, pp.16-26, 2017.

[72] A. Ben Tamou, L. Ballihi and D. Aboutajdine, "Automatic learning of articulated skeletons based on mean of 3d joints for efficient action recognition," in International Journal of Pattern Recognition and Artificial Intelligence, vol. 31, no. 04, 2017.

[73] A. A. Liu, W. Z. Nie, Y. T. Su, L. Ma, T. Hao and Z. X. Yang, "Coupled hidden conditional random fields for RGB-D human action recognition," in Signal Processing, vol. 112, pp. 74-82, 2015.

[74] Z. Liu, C. Zhang, Y. Tian, "3D-based deep convolutional neural network for action recognition with depth sequences," in Image Vis. Comput., vol. 55, pp. 93–100, 2015.

Kamal Kant Verma

Kamal Kant Verma is research scholar in Uttarakhand Technical University Dehradun. He is currently working as an Assistant Professor in Department of Computer Science & Engineering, COER Roorkee Uttarakhand India. He did B.Tech in Information Technology in 2006, M.Tech in CSE in 2012 and currently pursuing PhD from Uttarakhand Technical University Dehradun India. He has 15 years of teaching and research experience. His research area is Human Activity Recognition, Human Computer Interface, Pattern Recognition and Signal Processing. He has published more than 20 research papers in reputed national/ international journal and conferences such as Springer, Elsevier, IJIMAI, etc.

Brij Mohan Singh

Brij Mohan Singh is Director of College of Engineering Roorkee & Professor in Department of Computer Science and Engineering, COER Roorkee India. He has published more than 35 research papers in International Journals such as Document Analysis and Recognition-Springer, CSI Transactions on ICT-Springer, IJIG-World Scientific, IJMECS, EURASIP Journal on Image and Video Processing etc. His research areas are DIP and Pattern Recognition. He has guided 3 PhD Thesis of UTU and currently 6 are in process.

Acoustic Classification of Mosquitoes using Convolutional Neural Networks Combined with Activity Circadian Rhythm Information

Jaehoon Kim[1], Jeongkyu Oh[2], Tae-Young Heo[1]*

[1] Department of Information & Statistics, Chungbuk National University, Chungbuk (Republic of Korea)
[2] Data Scientist Team, BEGAS Inc., Seoul (Republic of Korea)

Received 1 June 2020 | Accepted 19 May 2021 | Published 9 August 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

Many researchers have used sound sensors to record audio data from insects, and used these data as inputs of machine learning algorithms to classify insect species. In image classification, the convolutional neural network (CNN), a well-known deep learning algorithm, achieves better performance than any other machine learning algorithm. This performance is affected by the characteristics of the convolution filter (ConvFilter) learned inside the network. Furthermore, CNN performs well in sound classification. Unlike image classification, however, there is little research on suitable ConvFilters for sound classification. Therefore, we compare the performances of three convolution filters, 1D-ConvFilter, 3×1 2D-ConvFilter, and 3×3 2D-ConvFilter, in two different network configurations, when classifying mosquitoes using audio data. In insect sound classification, most machine learning researchers use only audio data as input. However, a classification model, which combines other information such as activity circadian rhythm, should intuitively yield improved classification results. To utilize such relevant additional information, we propose a method that defines this information as a priori probabilities and combines them with CNN outputs. Of the networks, VGG13 with 3×3 2D-ConvFilter showed the best performance in classifying mosquito species, with an accuracy of 80.8%. Moreover, adding activity circadian rhythm information to the networks showed an average performance improvement of 5.5%. The VGG13 network with 1D-ConvFilter achieved the highest accuracy of 85.7% with the additional activity circadian rhythm information.

Keywords

Artificial Intelligence, Bayes' Rule, Convolutional Neural Network, Mosquitoes Classification, A Priori Probability.

DOI: 10.9781/ijimai.2021.08.009

I. Introduction

Mosquitoes are amongst the deadliest insects in the world and they have a direct impact on human lives. From malaria alone, 438,000 people died in 2015 [1]. In addition, Zika virus, Dengue, Chikungunya, and Yellow fever are all carried by *Aedes aegypti*, one of the most dangerous mosquito species. It is therefore not surprising that computational entomology, which records insect information and automatically classifies or detects pests, is studied more intensively than ever. Most computational entomology studies use insect image and sound data as important inputs to an algorithm. In an image classification study, Okayasu, Yoshida, Fuchida, and Nakamura. [2] photographed mosquitoes using a single-lens reflex (SLR) camera and mobile phone. The SLR camera images were used for learning in both conventional machine learning and deep learning algorithms. The performance of each algorithm was then tested using mobile phone images. Park, Kim, Choi, Kang, and Kwon [3] caught mosquitoes native to Korea and used their images as inputs for commercial deep learning algorithms such as visual geometry group (VGG), ResNet and SqueezeNet. The authors of [4] developed an inexpensive audio sensor and used it to classify *Bombus impatiens*, *Culex quinquefasciatus*, and *Aedes aegypti*. In [5], the authors obtained audio data from eight mosquitoes and two flies, and classified them.

Traditionally, machine learning algorithms for classifying an audio signal consist of three successive processes. First, the audio signal data for a certain period are converted into spectral-temporal parameters, including frequency and amplitude, which enables decomposition into components. In sound recognition, this spectral-temporal representation is generally used as input to a network. The performance of the network algorithms depends considerably on the type of spectral-temporal representation applied. Because Mel-frequency cepstral coefficient (MFCC) extracts information from human-recognized low frequencies, classifiers trained with this algorithm emulate human hearing. For this reason, most researchers use MFCC as a basic spectral-temporal representation [6]-[9]. In this study, we also utilize MFCC to provide input to classifiers. Second, to determine the signal classes, feature extraction transforms the MFCC data into descriptors representing each audio signal. Typical feature

* Corresponding author.

E-mail address: theo@cbnu.ac.kr

extraction methods include calculation of the average and standard deviation of spectral-temporal features, principal component analysis (PCA) [10], and Autoencoder [11]. Finally, conventional machine learning classifiers such as k-nearest neighbor (kNN), support vector machine (SVM), and random forest (RF) define data classes using the extracted descriptors as input.

However, a convolutional neural network (CNN) simultaneously performs feature extraction in the network to obtain a description. Unlike traditional feature extraction methods, such as PCA and Autoencoder, the convolution filter (ConvFilter) used in CNN learns with the goal of finding a precise description for the distinction of classes. In many studies, CNN is one of the highest performing algorithms in speech recognition as well as image classification. In the Rare Sound Event Detection Task of the IEEE Audio and Acoustic Signal Processing challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2017, the 1D convolutional recurrent neural network model recorded an F-score of 93.1 and error rate of 0.13 [12]. In the Acoustic Scene Classification challenge of DCASE 2019, various audio signals such as park, metro, and airport were used as inputs for machine learning algorithms. In this challenge, Naranjo-Alcazar, Perez-Castanos, Zuccarello, and Cobos. [13] compared prediction performances in relation to the number of layers used in a VGG-based network.

Since ConvFilter is responsible for feature extraction in CNN, the performance of CNN greatly depends on the type of filter and the method of layer stacking. For example, VGG, which won second place in the ImageNet Large Scale Visual Recognition challenge (ILSVRC) 2014, one of the most famous image classification challenges, used only a 3×3 ConvFilter. This network showed better performance than CNN models using different size filters [14]. In addition, GoogleNet, which won first place in ILSVRC 2014, was configured to reduce the computational demands of the model and calculate the correlation between channels using a 1×1 filter [15]. ResNet, which won first place in ILSVRC 2015, succeeded in building up to 152 layers using a skip connection that directly connected the layer input to its output [16]. In image classification, a CNN usually stacks layers using the 2D-ConvFilter. The 2D-ConvFilter is an intuitive filter in the image classification field, because it can create feature maps that detect high-level descriptions such as face, nose, and body from edge detection of the image. However, the spectral-temporal representation used as input in sound recognition requires a different approach in the use of ConvFilters, because this representation includes frequencies and amplitudes distributed over time, unlike images. Therefore, comparing the performances of ConvFilters for sound data is very important. Research has been conducted on the sound classification performance of commercial networks such as AlexNet, VGG, Inception, and ResNet [17]. Performance in electrocardiogram classification, which has the same form as audio signals, was studied using 1D-ConvFilter and 2D-ConvFilter [18]. In [19], the bulbul network using 3×1 2D-ConvFilter, and the sparrow network using 3×3 2D-ConvFilter were constructed for bird detection in audio signals. Sharma, Granmo, and Goodwin [20] used various spectral-temporal representations as inputs for their network. The proposed network obtains separate information for each representation, by stacking 3x1 2D-ConvFilter and 1×5 2D-ConvFilter. The accuracy of this network on the environmental sound classification dataset ESC-50 is 88.50%.

Biologically, a mosquito species' activity circadian rhythm refers to the probability of that species being active as a specific of time of day. In computational entomology, the activity circadian rhythm is significant information that can improve the performance of algorithms because different species have different activity cycles [5]. Given these activity cycles, if a trained classifier such as CNN, SVM, or RF is combined with activity cycle information, it will outperform uncombined classifiers. However, there are two major problems. The first problem is that most machine learning methods should retrain when significant new information is added. Second, the method of combination is not well established. Thus, when there is significant additional information, such as geographic distribution or activity circadian rhythm, we propose a simple method to define information as a *priori* probability, and combine it with a trained model using Bayes' rule [21]-[22]. Using this method, we avoid unnecessary relearning when new information is combined with a classification model. In addition, we can contribute to simplifying network learning using variables with different characteristics, such as activity circadian rhythms and audio signals.

The two main purposes of this paper are as follows: first, comparing the performances of 1D-ConvFilter, 3×1 2D-ConvFilter and 3×3 2D-ConvFilter, based on the VGG and a Simple CNN, to find a suitable network for mosquito classification using audio data. To train the networks, we use audio data, which include the signals of eight mosquitoes and two flies from [5]. The second purpose is to propose a simple method of combining the classification from audio signals with appropriate information of a different type. We demonstrate our proposed method by combining activity circadian rhythm information with our network classification. This simply requires the time of day to be recorded in the process of audio data collection.

II. Methods

A. Data

The data [23] provided by [5] are the wingbeat signals of eight mosquitoes and two flies obtained using audio sensors that are able to detect insects' wingbeats. The classes of the mosquitoes and the abbreviations of each class are listed in Table I. All the results of this study use these class abbreviations. According to [5], most of these insects were imported from different regions, such as California, Texas, and Taiwan, and were raised under specific conditions. A dataset of 50,000 samples, with 5,000 samples for each species, was built up.

A noise filter was applied to remove background noise from the wingbeat audio signal detected by the sensor. The audio signal was recorded at a sampling rate of 16 kHz and the duration of the signal was set to 1 s. In addition, where noise was removed, the position of the wingbeat signal was fixed to the center by the centering method, and zero-padding fixed the values in the remaining interval at 0.

The audio signal was converted into an MFCC with a shape of 40 × 43 × 1 to use as input for the models. The values 40, 43, and 1 denote the numbers of time, frequency, and amplitude intervals, respectively.

TABLE I. Abbreviations of Species

Abbreviation	Class
Fruit_Flies (FF)	Drosophila simulans
House_Flies (HF)	Musca domestica
Aedes_Female (AF)	Ae. aegypti (female)
Aedes_Male (AM)	Ae. aegypti (male)
Quinx_Female (QF)	Cx. quinquefasciatus (female)
Quinx_Male (QM)	Cx. quinquefasciatus (male)
Stigma_Female (SF)	Cx. stigmatosoma (female)
Stigma_Male (SM)	Cx. stigmatosoma (male)
Tarsalies_Female (TF)	Cx. tarsalis (female)
Tarsalies_Male (TM)	Cx. tarsalis (male)

B. Convolutional Neural Networks

The first purpose of this study was to compare the performances of three ConvFilters to determine the filter with the highest performance in classifying mosquitoes using audio data. For this comparison, we shared the same network structure with each of the filters. The ConvFilters used for classification were the 1D-ConvFilter, 3×1 2D-ConvFilter and 3×3 2D-ConvFilter, of size 3. Fig. 1 shows the feature extraction process for these three filters. 1D-ConvFilter extracts the description for a time domain of size 3 and the entire frequency domain. Additionally, 3×3 2D-ConvFilter extracts the local description of time × frequency as 3×3. However, 3×1 2D-ConvFilter obtains a separable description of a time domain for each frequency.

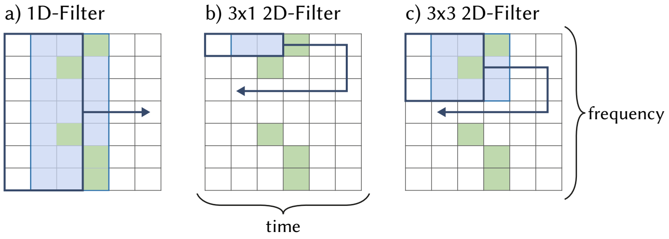


Fig. 1. Feature extraction process for three different filters.

Table II summarizes the simple CNN and VGG networks configured for filter comparison. Simple CNN is a model that measures the performance of each filter in a shallow network. This network has 6 layers, including the fully connected (FC) layer. The 1D-ConvFilter, 3×1 2D-ConvFilter, and 3×3 2D-ConvFilter have 0.6 million, 4.7 million, and 3.5 million parameters, respectively. In Simple CNN, the shapes of the final feature maps of the filters are 7×128, 7×10×128, and 7×7×128. Thus, the total number of parameters differs dramatically depending on the size of the feature map entering the FC layer.

TABLE II. Configuration Summaries of Convolutional Neural Networks

CNN	Filter	Input Shape	Number of Layers	Number of Parameters (M : million)
Simple CNN	1D-ConvFilter	40×43	6	0.6M
	3×1 2D-ConvFilter	40×43×1	6	4.7M
	3×3 2D-ConvFilter		6	3.5M
VGG13	1D-ConvFilter	40×43	13	22M
	3×1 2D-ConvFilter	40×43×1	13	22M
	3×3 2D-ConvFilter		13	28.3M

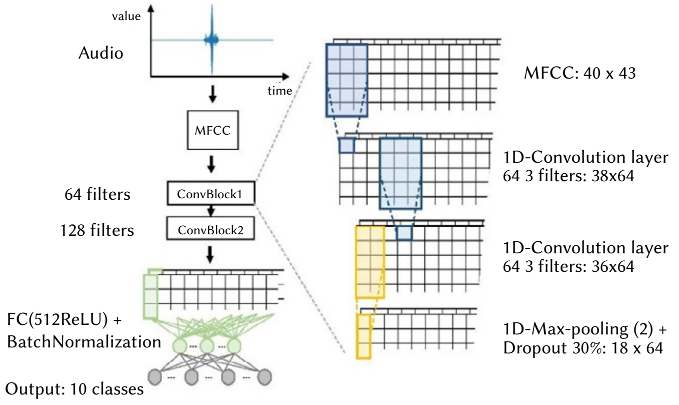


Fig. 2. Overall structure of the Simple CNN with 1D-ConvFilter (Simple 1D-CNN).

For VGG, we use VGG13 (configuration B of [14]) to adjust the shape of the feature map. To compare the performance of the ConvFilters, the filters of VGG13 were designated as 3×3 2D-ConvFilter, 3×1 2D-ConvFilter and 3 1D-ConvFilter. The numbers of parameters were approximately 22 million, 22 million, and 28.3 million. The shapes of the final feature maps of the filters were 1×512, 1×1×512, and 1×1×512. In VGG, the total number of parameters is determined by the different numbers of parameters of each filter, regardless of the size of the feature map entering the FC layer.

1. Simple CNN

We proposed three ConvFilters, with two shared CNN structures for each. The first shared network is Simple CNN with a shallow layer. To aid understanding, we describe a Simple CNN trained with 1D-ConvFilter.

Fig. 2 shows the overall structure of this Simple 1D-CNN. The layers consist of 1D-Convolution, 1D-Max-pooling, Dropout, BatchNormalization and FC. Each ConvBlock consists of two 1D-convolution layers of the same size, a 1D max-pooling layer with a kernel size of 2, and a dropout layer with a ratio of 30%. The kernel size of all ConvFilters is 3, and the 1D-convolution filter sizes of the two ConvBlocks are 64 and 128. For feature extraction, the first convolutional layer (ConvLayer) extracts a 38×64 feature map from a 40×43 shaped MFCC. The feature map provides descriptors of the entire frequency domain in a specific time domain MFCC, as previously described. The feature map channel is determined according to the number of filters. In total, the MFCC shrinks from 40×43 to 7×128 as it proceeds through the feature extraction. The FC layer, of size 512, uses descriptors obtained through the filters as inputs to classify the mosquito species. The activation function of the final FC layer, of size 10, uses softmax.

2. VGGNet

VGGNet is a commonly used CNN structure in many fields, because of the intuitiveness of its model structure. We now describe VGG13 with 3×3 2D-ConvFilter to illustrate how 2D-ConvFilter is learned inside a CNN. Essentially, 2D-ConvFilter moves in two dimensions in the MFCC and learns features locally. The deeper the layer, the more effective it is in creating high-level descriptors by combining local low-level descriptors.

Fig. 3 shows the overall structure of VGG13 with 3×3 2D-ConvFilter. VGG13 consists of 2D-Convolution, 2D max-pooling, dropout and FC layers. Each ConvBlock consists of two 2D-convolution layers with the same filter size, and a 2D max-pooling layer. The kernel size of all

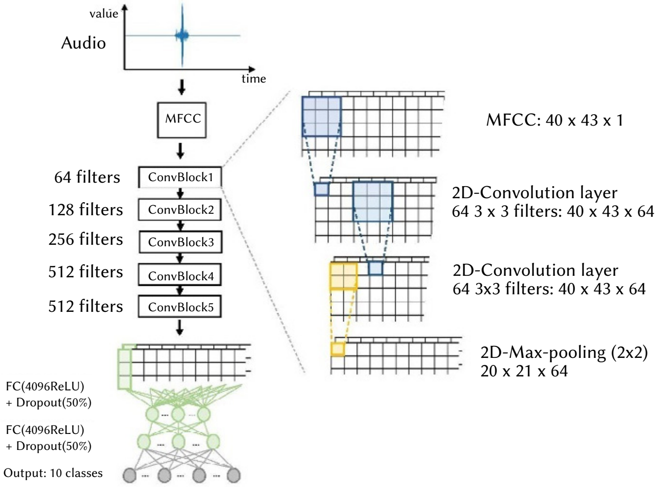


Fig. 3. Overall structure of VGG13 with 3×3 2D-ConvFilter.

filters is 3×3, and the filter sizes of the successive 2D-ConvBlocks are 64, 128, 256, 512 and 512**.** The shape of a feature map is interpreted as time × frequency × channels. In feature extraction, since all ConvLayers use padding, the first ConvLayer extracts a feature map of 40×43×64 from the 40×43×1 MFCC, which is the same size. Throughout feature extraction, MFCC is reduced from 40×43×1 to 1×1×512. The descriptors obtained through the filters classify the mosquito species after passing through two FC layers of size 4096 and a final FC layer of size 10. The activation function of the final layer uses softmax, as with Simple CNN.

C. Bayes' Rule-based Method for Adjusting Classification Output

In [5], a Naïve Bayes classifier is combined with activity cycle information, and it will outperform uncombined classifiers. However, most classifiers except the Naïve Bayes classifier face two major problems in combining activity cycle information. The first problem is that most machine learning methods should retrain when significant new information is added. Second, the method of the combination is not well established.

The appearance rate of insects differs according to information such as geographic distribution and activity circadian rhythm. Intuitively, if we have previously obtained information affecting the appearance rate, a trained classifier could use this information to obtain better performance. However, in most computational entomology studies, although such information regarding appearance rate is known, there is insufficient discussion about how to use it. In this study, we propose a method that defines prior information about appearance rate as a *priori* probability, and combines it with trained classifiers.

The prior information obtained by [5] was the activity cycle for each species. This was based on the time of observation of individual insects over one month. Fig. 4 shows the diurnal activity cycles, or activity circadian rhythms, of the ten species, identified by the abbreviations in Table I. In Fig. 4, there are two moments in the day in which there is a more notorious activity, of all the species in general. QM showed the most activity between 9 p.m. and 11 a.m., and TM showed the most activity between 5 a.m. and 7 a.m.

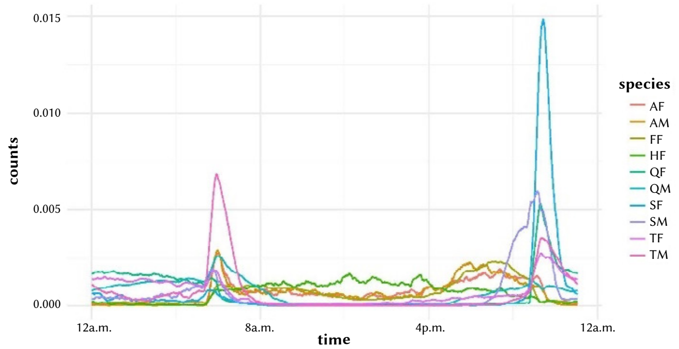


Fig. 4. Activity circadian rhythms for 10 species (see Table I for abbreviations).

The ultimate purpose of the adjusted classifier is to predict the new rate at which the i-th species c_i will appear when the independent variable, $\tilde{x}$, (here MFCC) exists at a certain time, t. We define an activity time rate as a *priori* probability $\hat{p}_{c_i}(t)$ for each species c_i and apply it to the trained CNN. Before calculating the predicted appearance rate of insect species, we assume that the scores of the training data $p(\tilde{x}|c_i)$ and the scores of the new data $\hat{p}_t(\tilde{x}|c_i)$ are the same.

Suppose we wish to predict the new appearance rate of a certain species at time. Bayes' rule provides

$$\hat{p}_t(c_i|\tilde{x}) = \frac{\hat{p}_t(\tilde{x}|c_i)\hat{p}_t(c_i)}{\hat{p}_t(\tilde{x})} \tag{1}$$

where a *priori* probability $\hat{p}_t(c_i)$ denotes the appearance rate of i-th species c_i at time t in the new data, and $\hat{p}_t(\tilde{x})$ denotes the marginal probability of $\hat{p}_t(\tilde{x}|c_i)$.

The estimated probability of the classifier for the training data $\hat{p}(c_i|\tilde{x})$ is as follows:

$$\hat{p}(c_i|\tilde{x}) = \frac{\hat{p}(\tilde{x}|c_i)\hat{p}(c_i)}{\hat{p}(\tilde{x})} \tag{2}$$

where a *priori* probability $\hat{p}(c_i)$ denotes the appearance rate of i-th species c_i in the training data, and $\hat{p}(\tilde{x})$ denotes the marginal probability of $\hat{p}(\tilde{x}|c_i)$.

Since the scores of the training data $\hat{p}(\tilde{x}|c_i)$ and the new data $\hat{p}_t(\tilde{x}|c_i)$ are the same, by equating equation (1) to (2) and defining $\mathrm{g}(\tilde{x}) = \hat{p}(\tilde{x})/\hat{p}_t(\tilde{x})$, we obtain

$$\hat{p}_t(c_i|\tilde{x}) = \mathrm{g}(\tilde{x})\frac{\hat{p}_t(c_i)}{\hat{p}(c_i)}\hat{p}(c_i|\tilde{x}) \tag{3}$$

Since $\sum_{i=1}^{n}\hat{p}_t(c_i|\tilde{x})=1$, we obtain $\mathrm{g}(\tilde{x})=[\sum_{i=1}^{n}\hat{p}_t(c_i)/\hat{p}(c_i)\cdot\hat{p}(c_i|\tilde{x})]^{-1}$. This also means that the term is statistically normalized.

Finally, by Bayes' rule, the relationship between the a *priori* probability $\hat{p}_t(c_i)$ in the above equation and the a *priori* probability for activity cycle $\hat{p}_{c_i}(t)$ is

$$\hat{p}_t(c_i) = \frac{\hat{p}_{c_i}(t)\hat{p}(c_i)}{\hat{p}(t)} \tag{4}$$

where the probability $\hat{p}(t)$ denotes the appearance rate at time t in the activity cycle. Since $\hat{p}_t(c_i|\tilde{x}) = \mathrm{g}(\tilde{x})\frac{\hat{p}_t(c_i)}{\hat{p}(c_i)}\hat{p}(c_i|\tilde{x})$, and $\hat{p}(t)$ is included in normalizing term $\mathrm{g}(\tilde{x})$, we can easily obtain $\hat{p}_t(c_i|\tilde{x}) \propto \hat{p}_{c_i}(t)\hat{p}(c_i|\tilde{x})$. The process of obtaining the *posteriori* probability above is illustrated in Fig. 5.

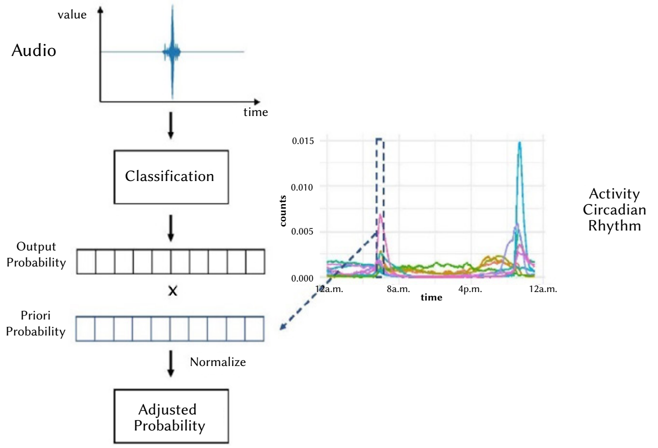


Fig. 5. Process of adjusting output probabilities using Bayes' rule method.

Suppose we have additional information such as geographic distribution as well as activity circadian rhythm, which together is expressed as m variables F_j, $j = 1 \ldots m$. If we assume that these variables are independent, and $\hat{p}(\tilde{x}|c_i) = \hat{p}_{Fs}(\tilde{x}|c_i)$, the a *posteriori* probability can be generalized as

$$\hat{p}_t(c_i|\tilde{x}) \propto \prod_{j=1}^{m}\hat{p}_{c_i}(F_j)\,\hat{p}(c_i|\tilde{x}) \tag{5}$$

D. Training the Convolutional Neural Networks

Training of neural networks involves the process of repeatedly adjusting weights to reduce differences between network-predicted and actual values to below a threshold. The tuning parameters required for network training are initializer, optimizer, epoch, and batch size. Simple CNN uses the Xavier initializer [24], which depends on the number of previous and next nodes. The VGG network uses the uniform initializer. Most sound classification models use the Adam optimizer [19]-[20]. We also use the Adam optimizer [25], in which decay rates $\beta_1\beta_1$ and $\beta_2\beta_2$ are set to 0.9 and 0.999, respectively, and the learning rates are set to 1e-5 and 1e-4 for optimal network selection. The epoch of the network is fixed at 100, and the batch sizes are set at 128 and 256. To train the networks, we used the Keras framework in a Ryzen 2700x @3.70 GHz with 32 GB RAM and RTX 2080ti.

E. K-fold Cross-Validation

The k-fold cross-validation method divides the training dataset into k data subsets. Next, one of the k subsets is used for model evaluation, and the remaining k-1 subsets are used as training data. By repeating this process k times, k-fold cross-validation uses all the subsets as validation data. The final accuracy of the classifier is the average of the k-fold accuracies.

In this study, we used 5-fold cross-validation to compare the performances of the networks. In addition, we separated the data into 80% training-set and 20% validation-set, for comparison of networks that had been trained only with MFCC and those that were combined with activity circadian rhythm information.

III. Results

A. Classification Performance

Our metric for evaluating classification performance is the accuracy of each network obtained by 5-fold cross-validation. For each model, learning rates of 1e-5 and 1e-4 were applied, and batch sizes of 256 and 128 were used. Table III shows the average accuracies of Simple CNNs using 5-fold cross-validation. For Simple CNNs with 4 ConvLayers, the 1D-ConvFilter with learning rate 1e-4 and batch size 128 has an accuracy of 80.0%, higher than any of the 2D-ConvFilter configurations.

Table IV shows the average accuracies of VGG13 networks, with the same layout as Table III. Here, the highest accuracy of 80.8% is obtained for the VGG13 with 3×3 2D-ConvFilter, a learning rate of 1e-5 and a batch size of 256.

TABLE III. Comparison of Networks for Simple CNNs with Different Convolution Filters Using 5-Fold Cross Validation

CNN	Filter	Learning Rate	Batch Size	Accuracy
Simple CNN	1D-ConvFilter	1e-5	256	78.4%
		1e-5	128	79.3%
		1e-4	128	**80.0%**
	3×1 2D-ConvFilter	1e-5	256	77.4%
		1e-5	128	78.3%
		1e-4	128	79.1%
	3×3 2D-ConvFilter	1e-5	256	78.9%
		1e-5	128	79.8%
		1e-4	128	79.8%

TABLE IV. Comparison of Networks for Vgg13 with Different Convolution Filters Using 5-Fold Cross Validation

CNN	Filter	Learning Rate	Batch Size	Accuracy
Simple CNN	1D-ConvFilter	1e-5	256	80.5%
		1e-5	128	80.1%
		1e-4	128	79.0%
	3×1 2D-ConvFilter	1e-5	256	80.3%
		1e-5	128	80.4%
		1e-4	128	79.9%
	3×3 2D-ConvFilter	1e-5	256	**80.8%**
		1e-5	128	80.6%
		1e-4	128	80.1%

We conclude from Table III and Table IV that the 1D-ConvFilter shows the highest performance when the number of network layers is small and the 2D-ConvFilter shows the highest performance when the network is deeper.

B. Effect of Activity Circadian Rhythms on A Priori Probabilities

In Section II, we described CNNs with different ConvFilters, and explained how combining a trained network with significant a *priori* information could be used to obtain improved predictions. In this section, we discuss the effect of using activity circadian rhythms as a *priori* information to aid mosquito species classification. To this end, 50,000 audio datasets were divided into an 80% training-set and a 20% test-set. The learning rate and epoch for network training were set to 1e-5 and 256, respectively. The values of the remaining tuning parameters were the same as in the previous results.

Before discussing the results, we first describe the nature of the Naive Bayes classifier. Because we use Bayes' rule to train the classifier, a Naive Bayes classifier is more flexible than other classifiers for the problem of applying additional information. In [5], the Naive Bayes classifier was trained to use insect sound data and activity cycle information as input to the classifier. Table V shows the accuracy of each network according to whether or not activity cycle information was added. The average difference between networks with and without activity cycle information is approximately 5.5%. In addition, all VGG13 networks have higher accuracy than the reference accuracy of [5]. Generally, the networks without activity cycles have similar results to Table III and Table IV. Moreover, when applying activity cycles as additional information, the accuracy of the 1D-ConvFilter in Simple CNN is highest at 84.68%. However, unlike the results in Table IV, the accuracy of the 1D-ConvFilter in VGG13 is highest at 85.72%.

Fig. 6 shows the change in recall of VGG13 when activity circadian rhythm is added. Recall represents the ratio of the predicted number in the *i*-th class to the actual number in the *i*-th class. In other words, this measure indicates how well the classifier predicts the mosquito species for each sound signal. In Fig. 6, the overall recall of QF is noticeably lower than other classes. Conversely, AM has the highest average recall. We see that a network with activity cycle information (circle in Fig. 6) has a higher recall of all classes by about 1% difference than a network without this information (triangle). This improvement in recall is largest for AF, and smallest for FF. In addition, in the result of 1D-ConvFilter VGG13 with activity circadian rhythm, the average difference in recall for each class is significantly higher than the differences for other networks. This result causes the highest accuracy 85.72% for 1D-ConvFilter in Table IV results for VGG13.

TABLE V. Comparison of Networks with or without Addition of Activity Circadian Rhythms

Adding Activity Circadian Rhythms	CNN	Filter	Accuracy
No	Simple CNN	1D-ConvFilter	78.76%
		3×1 2D-ConvFilter	76.46%
		3×3 2D-ConvFilter	74.90%
	VGG13	1D-ConvFilter	80.36%
		3×1 2D-ConvFilter	80.41%
		3×3 2D-ConvFilter	80.47%
Yes	Simple CNN	1D-ConvFilter	**84.68%**
		3×1 2D-ConvFilter	83.40%
		3×3 2D-ConvFilter	82.28%
	VGG13	1D-ConvFilter	**85.72%**
		3×1 2D-ConvFilter	85.66%
		3×3 2D-ConvFilter	82.91%
	Naive Bayes Method [5]		79.44%

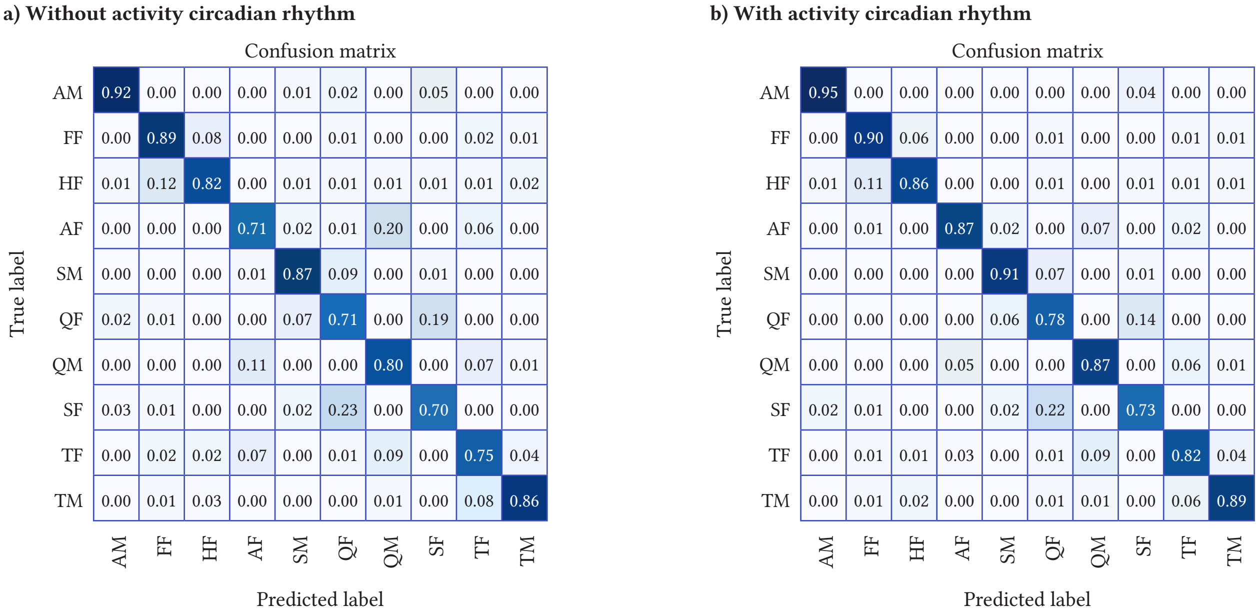


Fig. 7. Confusion matrix of VGG13 with 3×3 2D-ConvFilter (a, left) without and (b, right) with activity circadian rhythm information added.

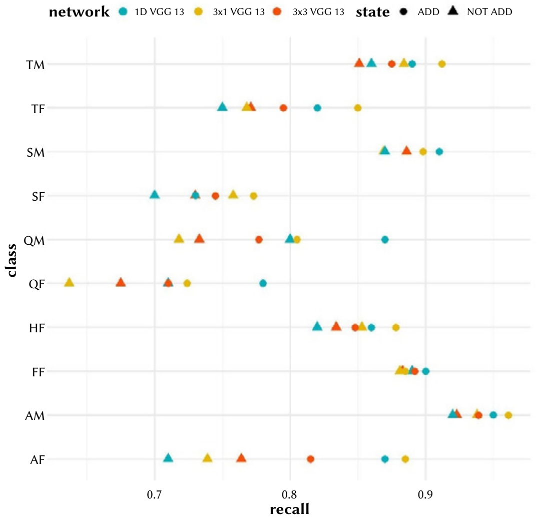


Fig. 6. Scatter plot of recall of different VGG13 networks without (triangles) and with (circles) activity cycle information.

Fig. 7 shows the change in the confusion matrix of VGG13 using 1D-ConvFilter from (a) without to (b) with activity circadian rhythm information. This network gave the highest accuracy with activity information (Table V). Without activity circadian rhythm, the AF to QM and QM to AF prediction rates account for 20% and 11% of misclassifications, respectively. However, with activity circadian rhythm, these rates fall to 7% and 5%. Other cases similarly show that adding activity circadian rhythm information reduces misclassification errors.

IV. Discussion and Limitation

In this section, we discuss some contributions and limitations of our study. Traditionally, CNN is a method in which research on image data is becoming active. And ConvFilter, which determines the performance of this method, is also being studied a lot on image data. However, research on ConvFilter is not active on sound data. So we introduced ConvFilters in section Ⅱ and analyzed their results in section Ⅲ. Simple CNN with the 1D-ConvFilter has an accuracy of 80.0%, higher than any of the Simple CNNs with 2D-ConvFilter. In VGG13, the highest accuracy of 80.8% is obtained for the VGG13 with 3×3 2D-ConvFilter.

When applying activity cycles as additional information, the accuracy of Naive Bayes classifier [5] is 79.44%. On the other hand, the accuracies of our proposed methods are 84.68% and 85.72% at the 1D-ConvFilter in Simple CNN and the 1D-ConvFilter in VGG13, respectively. Moreover, adding activity circadian rhythm information to each of the VGG 13 and Simple CNNs results in a 5.5% and 6.7% difference in average performance improvement, respectively. In Fig. 6, the recall values and misclassification rates of each class show better results by about 10% difference with adding activity circadian rhythm information. Thus, it is explained that additional information such as activity rhythm information improves the performance of the network.

While our evaluations are encouraging, there are certain limitations to our method. We proceed with the analysis using limited data. If the data containing other information such as location and seasonality as well as activity circadian rhythm information are used, the analysis results are more reliable. In order to extract features of sound data, feature extraction methods other than MFCC may be used. Furthermore, in order to compare with CNN models, we can apply end-to-end neural network models that take sound data of mosquitoes as input.

V. Conclusion

The first objective of this study was to find a network filter configuration that could use audio signals to classify mosquitoes and flies. We selected three different filters, 1D-ConvFilter, 3×1 2D-ConvFilter, and 3×3 2D-ConvFilter, and applied them to Simple CNN and VGG13 networks to classify mosquito and fly species. The accuracy of each network was calculated using 5-fold cross-validation. Comparing the results, VGG13 with 3×3 2D-ConvFilter showed the highest accuracy of 80.8%. Also, all the accuracies of VGG13 networks are greater than that of Simple CNN.

Second, because different species have different activity cycles, we proposed a method using Bayes' rule to combine activity cycle information with trained networks. The activity circadian rhythm for each species was defined as an a *priori* probability to use Bayes' rule.

The adjusted probability for each species was obtained by multiplying the defined a *priori* probability by the probability obtained from the trained network. Combining networks with activity cycles in this way showed an average improvement in accuracy of 5.5%, with VGG13 using 1D-ConvFilter showing the highest accuracy of 85.72%. Furthermore, by incorporating activity cycle information, misclassifications, such as AF to QM, can be reduced.

In conclusion, when performing classification, we can use not only audio data or image data, but also other types of information, such as activity cycle and geographical distribution. Thus, if location and time information are also collected in the process of collecting audio data, we believe that this relatively simple method can obtain even better results.

References

[1] World Health Organization: Mosquito-Borne Diseases. Available online: http://www.who.int/neglected_diseases/vector_ecology/mosquito-borne-diseases/en/ (accessed on 1 April 2017).

[2] K. Okayasu, K. Yoshida, M. Fuchida, and A. Nakamura, "Vision-Based Classification of Mosquito Species: Comparison of Conventional and Deep Learning Methods," *Applied Sciences*, vol. 10, no. 1, pp. 3935, 2019, doi:10.3390/app9183935.

[3] J. Park, D. I. Kim, B. Choi, W. Kang, and H. W. Kwon, "Classification and Morphological Analysis of Vector Mosquitoes using Deep Convolutional Neural Networks," *Scientific Reports*, vol. 10, no. 1, pp. 1-12, 2020, doi: 10.1038/s41598-020-57875-1.

[4] G. E. Batista, E. J. Keogh, A. Mafra-Neto, and E. Rowton, "SIGKDD demo: sensors and software to allow computational entomology, an emerging application of data mining," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2011, pp. 761-764, doi: 10.1145/2020408.2020530.

[5] Y. Chen, A. Why, G. E. Batista, A. Mafra-Neto, and E. J. Keogh, "Flying insect classification with inexpensive sensors," *Journal of insect behavior*, vol. 27, no.5, pp. 657-677, 2014, doi: 10.1007/s10905-014-9454-4.

[6] J. J. Noda, C. M. Travieso-González, D. Sánchez-Rodríguez, and J. B. Alonso-Hernández, "Acoustic Classification of Singing Insects Based on MFCC/LFCC Fusion," *Applied Sciences*, vol. 9, no. 19, pp. 4097, 2019, doi: 10.3390/app9194097.

[7] Z. Le-Qing, "Insect sound recognition based on MFCC and PNN," in *2011 International Conference on Multimedia and Signal Processing*, 2011, pp. 42-46, doi: 10.1109/CMSP.2011.100.

[8] N. Saleem, and T. G. Tareen, "Spectral Restoration based speech enhancement for robust speaker identification," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 1, pp. 34-39, 2018, doi: 10.9781/ijimai.2018.01.002.

[9] N. Saleem, and M. I. Khattak, "Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 84-90, 2020, doi: 10.9781/ijimai.2019.06.001.

[10] X. Fan, H. Feng, and M. Yuan, "PCA based on mutual information for acoustic environment classification," In *2012 International Conference on Audio, Language and Image Processing*, 2012, pp. 270-275, doi: 10.1109/ICALIP.2012.6376624.

[11] S. Ghosh, E. Laksana, L. P. Morency, and S. Scherer, "Learning representations of affect from speech," 2015, arXiv:1511.04747.

[12] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 80-84.

[13] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "DCASE 2019: CNN depth analysis with different channel inputs for Acoustic Scene Classification," 2019, *arXiv:1906.04591.*

[14] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556.*

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[17] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE international conference on acoustics, speech and signal processing*, 2017, pp. 131-135, doi: 10.1109/ICASSP.2017.7952132.

[18] Y. Wu, F. Yang, Y. Liu, X. Zha, and S. Yuan. "A comparison of 1-D and 2-D deep convolutional neural networks in ECG classification," 2018, *arXiv:1810.07088.*

[19] T. Grill, and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1764-1768, doi: 10.23919/EUSIPCO.2017.8081512.

[20] J. Sharma, O. C. Granmo, and M. Goodwin, "Environment Sound Classification using Multiple Feature Channels and Attention based Deep Convolutional Neural Network," 2019, *arXiv:1908.11219.*

[21] M. Saerens, P. Latinne, and C. Decaestecker, "Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure," *Neural computation*, vol. 14, no. 1, pp. 21-41, 2002, doi: 10.1162/089976602753284446.

[22] P. Latinne, M. Saerens, and C. Decaestecker, "Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: evidence from a multi-class problem in remote sensing," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, Vol. 1, pp. 298-305.

[23] Y. Chen, A. Why, G. E. Batista, A. Mafra-Neto, and E. J. Keogh. "Flying Insect Classification with Inexpensive Sensors." Distributed by Y. Chen. https://sites.google.com/site/insectclassification/.

[24] X. Glorot, and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249-256.

[25] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980.*

Jaehoon Kim

Jaehoon Kim received the B.S. and M.S. in Information & Statistic from Chungbuk National University, South Korea. Now he is Ph. D course in statistics from Chungbuk National University. His research interests are hyperparameter optimization problem, computer vision, digital signal processing, data mining, statistical learning, and deep learning.

Jeonkyu Oh

He obtained his B.S and M.S. in Statistic from Chungbuk National University, South Korea. Now he is working as data scientist in BEGAS Inc. in South Korea. His area of interest includes machine learning, data mining and statistical learning.

Tae-Young Heo

He received the B. S. in Statistics from the Chungbuk National University and the M. S. and Ph. D. in Statistics from the North Carolina State University. Now he is a professor in department of Information and Statistics, Chungbuk National University. His current research interests include statistical learning and modeling.

Deep Feature Representation and Similarity Matrix based Noise Label Refinement Method for Efficient Face Annotation

A. Suruliandi[1], A. Kasthuri[1], S.P. Raja[2] *

[1] Department of Computer Science & Engineering, Manonmaniam Sundaranar University, Tirunelveli 627012 (India)
[2] Department of Computer Science & Engineering, Vellore Institute of Technology, Vellore, Tamilnadu (India)

Received 28 August 2020 | Accepted 19 March 2021 | Published 3 May 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

Face annotation is a naming procedure that assigns the correct name to a person emerging from an image. Faces that are manually annotated by people in online applications include incorrect labels, giving rise to the issue of label ambiguity. This may lead to mislabelling in face annotation. Consequently, an efficient method is still essential to enhance the reliability of face annotation. Hence, in this work, a novel method named the Similarity Matrix-based Noise Label Refinement (SMNLR) is proposed, which effectively predicts the accurate label from the noisy labelled facial images. To enhance the performance of the proposed method, the deep learning technique named Convolutional Neural Networks (CNN) is used for feature representation. Several experiments are conducted to evaluate the effectiveness of the proposed face annotation method using the LFW, IMFDB and Yahoo datasets. The experimental results clearly illustrate the robustness of the proposed SMNLR method in dealing with noisy labelled faces.

Keywords

CNN, Deep Learning, Face Annotation, Noise Label Refinement, Similarity Matrix.

DOI: 10.9781/ijimai.2021.05.001

I. Introduction

Recent years have witnessed the rapid growth of digital cameras and mobile devices, powerful cloud computing facilities, Web 2.0 photo sharing portals and social networks. Social media repositories such as Facebook, Twitter, Flickr, YouTube and Picasa allow users to upload and share personal photos or videos. As a consequence, masses of images have been created, distributed and shared on the internet by millions of users today, resulting in a large quantum of image collections on online social networks. Consequently, image sharing sites have difficulty managing and retrieving huge aggregates of face images. The plethora of multimedia content accessible today demands that challenges in terms of its storage, organization and indexing for future search and access be addressed. Moreover, an important aspect of online social media services is that users can annotate face images with keywords called tags, labels or captions. This voluntary activity of users who annotate faces with labels is termed labelling. Such labels may, however, be incorrect, imprecise or incomplete. Studies [1]-[3] show that name labels provided by users are highly "noisy", in the sense that only around 50% are actually appropriate to the corresponding person, because there are no restrictions or boundaries on assigning names to images on social media applications.

Due to the noisy nature of web facial images, early name labels of such web facial image databases were perhaps imperfect or damaged, in the absence of additional manual fine-tuning endeavours. A key technique that addresses this challenge is auto face annotation, which automatically assigns a name to the face of the corresponding person. Making an annotation reliable under noisy labeled facial images is a major challenge for real-life face annotation systems. To facilitate noise label refining and annotating huge facial image databases, several automatic face annotation methods have been proposed in the related work [4]-[9]. However, the labelling results reported fall short of the standards required of existing, reliable face annotation systems, especially in terms of real-time issues and noisy labels. Facial images normally have issues with variations in appearance, pose, illumination, occlusion, and noisy labels, all of which can result in mislabeling in face annotation. An efficient face annotation method must overcome these complications with innovative image mining abilities that capture discriminative and intrinsic information in faces. Moreover, sophisticated noise label refining capabilities are required to make the face annotation method robust. Hence, this paper proposes a new face annotation method, Similarity Matrix-based Noise Label Refinement (SMNLR), which concurrently deals with the problems of refining noise labels and assigning labels to facial images.

The face annotation method based on distance metric learning refines noisy labels powerfully and enhances the reliability of face

* Corresponding author.

E-mail addresses: suruliandi@yahoo.com (A. Suruliandi), kasthurianburajan@gmail.com (A. Kasthuri), avemariaraja@gmail.com (S.P.Raja).

annotation. The use of distance metric learning methods also implies that the appearance of facial features is not identical. Essentially, these methods are most appropriate for high-level noisy labels, and enhance the accuracy of face annotation. Thus, the proposed method refines human-provided unreliable labels by dropping inappropriate labels and adding missing ones. Additionally, the proposed method generates a suggested name list based on visual similarities for better face naming.

Generally, feature extraction techniques play a vital role in large collections of facial images by annotating them. Most of the existing face annotation methods [10]-[14] utilize the hand-crafted features for feature representation. Given that hand-crafted features are not adequate enough to handle the task, face annotation needs different levels of detailed descriptions to distinguish between faces in multi-granularity similarities. To tackle this problem, deep features are extracted from the deep network to describe face images. Deep networks, such as Convolutional Neural Networks (CNN) [15]-[18], offer superior multilevel facial representation. The CNN provides the highest number of descriptive features and is the least sensitive to real-time challenges. Recent researches [19]-[25] on facial image analysis state that deep features are more robust for such complex tasks. Hence, in this work, a CNN model is used for deep feature extraction. This CNN can effectively provide deep features from the face image and significantly improve annotation performance. The main contributions of this paper are, 1) A modified CNN architecture is introduced for deep feature extraction 2) A Similarity Matrix-based Noise Label Refinement (SMNLR) method is proposed to handle noisy labeled face images in a large-scale dataset. Inconsistent name labels can be effectively discovered by the probabilities of similarity measurements, and then fine-tuned or relabeled for training 3) The modified CNN with a proposed SMNLR method obtains state-of-the-art results on various face datasets, i.e., LFW, Yahoo, and IMFDB datasets.

A. Related Works

In recent years, Convolutional Neural Networks (CNNs) have shown an extraordinary ability for face feature representation in face annotation tasks. Several works [15], [26]-[28] on face applications indicate that deep feature extraction is more robust for such complex tasks. Ma et al. [29] combined the CNN model, AlexNet, with the proposed semantic extension model (SEM). CNN feature are provided as input for the proposed model. Problems with image tag refinement and assignment are overcome by using a self-defined Bayesian-based model which divides images with similar features into a semantic neighbor group. Venkatesh et al. [20] proposed the canonical correlation analysis (CCA) framework to facilitate a CNN feature and word-embedding vector. The CCA-KNN outperforms the Corel-5k, ESP-Game and IAPRTC-12 datasets. De Souza et al. [15] integrated the LBP feature descriptor with a modified Convolutional Neural Network (CNN) and proposed a new deep neural network called the LBPnet. An extended version of the LBPnet, called n-LBPnet, is also proposed. This method extracts deep features and outperforms other state-of-the-art techniques on the spoofing database. Kurban et al. [30] used the Eurecom Kinect Face dataset and Body Login Gesture Silhouettes dataset to create a virtual dataset of multimodal biometrics. Their study proves that Convolutional Neural Network (CNN)-based methods get better features and are also less sensitive to variations in pose, lighting and facial expressions in images.

Zeng et al. [31] have proposed a novel framework called Partial Permutation Matrix (PPM) for each image. In PPM, the samples of the same class from each image are related diagonally to the image set. SVM been introduced for labeling face images with names. Cour et al. [32] proposed a convex learning formulation based on minimizing a loss function suitable for partial label setting. The aim is to learn a classifier that can disambiguate partially labeled and ambiguously labeled images. Chen et al. [3] proposed a matrix completion for ambiguity resolution (MCAR) technique to calculate exact labels from unclearly labeled images. Noisy soft labeling vectors can, however, impact its performance. Consequently, iterative candidate elimination (ICE) procedure is applied to reduce the iterative ambiguity resolution by slowly eliminating parts of a vaguely labeled face. Liu et al. [33] proposed a self-error-correcting CNN (SECCNN) approach to work with noisy labels. The SECCNN develops a confidence policy that switches between the label of the sample and the max-activated output neuron of the CNN. Su et al. [34] have identified the difficulty of relating names with faces from large scale news images with captions. This problem was overcome by Person-based Subset Clustering which is mainly based on face clustering. This method provides the visual structural information all face images derived from the same name. Kumar et al. [35] proposed a two-step approach for both detection and recognition tasks. In the first step, a seed set is generated from the given image collection using detection and recognition algorithms. In the second step, the performance is improved by adapting the seed set. Maihani et al. [36] proposed a novel method for automatic image annotation wherein similar images are retrieved and a relative graph generated with tags. Finally, the tags of the dense community are chosen for the query image. Wang et al. [6] introduced an unsupervised label refinement (ULR) method to fine-tune weak labelled face images on online social networks. Their work uses a cluster-based approximation scheme for label refinement, while the majority voting approach is applied to tag names with facial images. The drawback of the ULR is that it cannot handle issues with duplicate names in real-life environments. Zhu et al. [8] proposed a knowledge transfer framework for face photo-sketch synthesis task. A new network architecture which allows to transfer knowledge from two teacher models to two student models are trained and knowledge has been transferred between two student models mutually. Two students network are trained using a small set of photo sketch pairs. Experimental results demonstrate that their proposed method performs better than other state-of-the-art methods. Zhu et al. [37] proposed a deep Convolutional Neural Network, to represent face photos. More precise person sketch patches and weight combination for sketch patch reconstruction could be obtained from the deep feature representations. Deep feature model based on the graphical representation is proposed to mutually discover weights for deep feature representations and reconstruction weights. Zhu et al. [38] proposed a deep collaborative framework with two opposite networks. These two networks perform the common communication between two opposite mappings. A collaborative loss is proposed in this work to limit the two contrary mappings and create them more balanced, as a result building the models more appropriate for photo–sketch synthesis task. Wang et al. [39] proposed a novel co-mining framework that utilizes two peer networks to identify the noisy faces, replaces the high-confidence clean faces and reassigns the clean faces in a mini-batch fashion.

B. Motivation and Justification

Most of the existing methods [4], [10], [40], [41] are applied directly on labeled facial images for face annotation without fine-tuning the labels, culminating in noisy or incorrect labels in face-name association. Certain early studies [1], [6], [42] overcame this drawback using unsupervised clustering algorithms to refine noise labels. In these clustering algorithms, a face collection is divided into several groups based on the identity name. Noisy labels are refined by estimating the maximal cluster among the groups of faces. However, the algorithms cannot prove that a face image indisputably belongs to a particular identity name; rather, they simply state that there is a high probability of the face image corresponding to the identity in question. This kind of simple correlation between faces and labels is not effective enough to refine label ambiguity. Consequently, several researchers [3], [43] have attempted to resolve the incompatibility between faces and name labels with supervised distance metric learning approaches. Distance metric learning-based label refinement techniques have shown better results than other existing label refinement techniques. In complex cases, however, information transmission follows no standard form and varies in feature gaps, a drawback that limits face annotation. Therefore, a much more accurate and robust noise label refinement technique is essential for effective face annotation by refining noise from labeled facial images. Thus motivated, an effort is made in this work to address the issue, and a new distance metric learning-based noise label refinement method is proposed, called the Similarity Matrix-based Noise Label Refinement (SMNLR), it combines the Cosine and Mahalanobis distance measures.

At the same time, the number of variations in faces also gives rise to the issue of label ambiguity because facial images are generally captured under various issues such as illumination, occlusion, expressions, and variations in poses. Most of the existing face annotation methods [1], [2] consider only hand-crafted feature extraction techniques for feature representation. They effectively capture the most information from facial images, and try to resolve issues by using a single or double layer to extract facial features. But, in several difficult domains, such as twin persons, these hand-crafted features generate the similar features for different persons due to its limitations. Hence, the faces might attain association with irrelevant labels in the context of label refinement owing to the low quality facial features. Also, when it deals with misaligned faces, it generates the unwanted texture information of faces. Hence, a robust feature is to be extracted from face images by overcoming these issues to improve the reliability of proposed face annotation method. Instead of utilizing the hand-crafted features, in recent years, Convolutional Neural Networks (CNN) [15], [44] extracts facial features using multiple levels of layers, wherein every single layer extracts deep features from faces. The CNN's remarkable learning features have helped resolve a variety of computer vision problems. These include image annotation, face recognition, image classification, object detection and identification, indicating that using deep features in face annotation for feature representation would be most efficient. Therefore, in this work, a most effective deep feature is used for feature representation in proposed SMNLR method.

The proposed SMNLR method effectively explores noisy labels by utilizing a fusion of the two discriminative similarity matrices. From the point of view of the literature, it is observed that the Cosine [45], [46] and Mahalanobis [47], [48] distance metric learning methods represent the most powerful similarity information between faces, compared to other existing distance metric learning approaches. The Cosine distance metric provides the direction information between samples, based on a broad collection of orientations. The Cosine of the orientation has essential uniform information for the matching components of faces. However, it does not consider magnitude differences between samples. Consequently, in critical circumstances involving illumination and expression, the cosine distance metric is too complex to handle all of the matching similarity information in the samples. To overcome this shortcoming, the Mahalanobis distance metric activates the similarity matrix by incorporating the magnitude difference of the relationship between the samples. Generally, the Mahalanobis distance metric encodes more meaningful similarity measurements using the uniform distribution of the sample with respect to face reconstruction. Therefore, this work combines the direction-based cosine similarity matrix and the distribution-based Mahalanobis similarity matrix. Therefore, this work combines the direction-based Cosine similarity matrix and the distribution-based Mahalanobis similarity matrix. Since the fusion of the two discriminative matrices uses a normalization parameter, α, with a value of 0.5, it significantly eliminates noisy labels and reassigns correct labels, based on the distance of the least similarity value of the fused similarity matrix. Justified by this, a new distance metric learning-based face annotation method called the SMNLR is proposed to refine noise labels based on a fusion of the Cosine and Mahalanobis similarity matrices. In addition, when the corresponding test face is not found in the training dataset, the given test face image is annotated with a name, using the suggested labels list. The suggested name list contains a list of labels that are applied when the test face does not match with database images of the training set. Given the need to name unknown faces in the test image, the list of suggested names is considered. The procedure for creating a suggested list further enhances the reliability of the proposed SMNLR method.

C. Outline of the Proposed Work

The outline of the face annotation process using the proposed SMNLR method is described in Fig. 1. The method comprises two phases, training and testing. The appropriate face region is chosen from the images to remove irrelevant information in the pre-processing step. In the training phase, deep features are extracted from the training images using the CNN. Two discriminative similarity matrices, the Cosine and Mahalanobis, are obtained using the training features and combined to create a fused similarity matrix. Noisy labels are refined and unambiguous labels reassigned, based on the similarity measurement of the fused similarity matrix. A suggested name list is also generated for face naming. In the testing phase, just as in the training phase, a feature extraction procedure is considered. The multi-class SVM classifier annotates the face images with their names.

D. Organization of the Paper

Section II explains the proposed SMNLR method in detail. Section III describes the databases and experimental results. Section IV discusses the performance analysis of the proposed method. Section V concludes the paper.

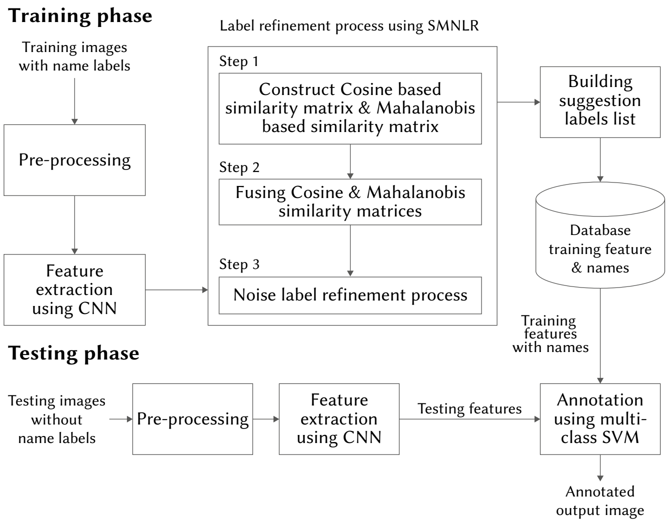

Training phase
Training images with name labels
Pre-processing
Feature extraction using CNN
Label refinement process using SMNLR
Step 1
Construct Cosine based similarity matrix & Mahalanobis based similarity matrix
Step 2
Fusing Cosine & Mahalanobis similarity matrices
Step 3
Noise label refinement process
Building suggestion labels list
Database training feature & names
Training features with names
Testing phase
Testing images without name labels
Pre-processing
Feature extraction using CNN
Testing features
Annotation using multi-class SVM
Annotated output image

Fig. 1. Process flow of the proposed SMNLR method.

II. The Proposed Method

A. Convolutional Neural Networks (CNN) Feature Extraction

Deep networks, such as Convolutional Neural Networks (CNN) [49], offer superior multilevel facial representation. The CNN model uses the output of a layer in the centre of the model as another description of the data, it is represented as a deep feature. Generally, CNN architecture consists of one or more convolution layers, often with a pooling layer, which are followed by one or more fully connected layers as in a neural network. The CNN uses this architecture to efficiently extract essential features from face image.

In this work, the CNN architecture consists of input layer, three convolutional layers, namely, convolution 1, convolution 2, and convolution 3; and three pooling layers, namely, pooling 1, pooling 2, and pooling 3. The input layer assigns an input image to the first convolutional layer. Convolutional layers play a major role in the CNN for feature extraction. Several convolutions can be performed on an input image, each utilizing a different filter and producing a unique feature map. Hence, the output of each layer describes a particular feature representation obtained from the input image. The convolution layer parameters contain several spatial and spectral learnable kernels or filters. In the first, second, and third convolutional layers, several feature maps are generated. Each convolutional layer is connected with a rectified linear unit (ReLU) and a pooling layer (down-sampling). A ReLU is an extensively used nonlinear activation function and presents a threshold operation to every component of the feature map. It assigns a value of 0 to the negative elements of the feature map. Pooling layers reduce the large number of features generated by the convolution layer. The convolved feature map is rendered more powerful and robust through the pooling layer. Max and average pooling are the most widely used techniques for the pooling process. The max pooling process is carried out by selecting the highest value of all the pixels in the receptive field to describe the output of the pooling feature map. The pooled (down-sampled) features generated in each pooling layer are provided as input to the next convolutional layer. Dropout layer is used to avoid the overfitting problem of features. Finally, fully connected layer generates deep feature values by combining all of the features learned from previous layers. A comprehensive demonstration of the CNN is shown in Fig. 2.

B. Convolution Layers

In convolution 1, 4 convolution filters with size of 4 × 4 are applied for the convolution process to generate feature maps. The convolution filter is applied with the stride of 1 to the input image. The convolution process is performed using Equation (1).

$$FM_p(x) = \sum\nolimits_{\forall y \epsilon N(x)} G(y) * K_p(m) \tag{1}$$

where FM_p (x) is the output feature map of the convolution process, where p = 1, 2, ..., 4 represents the p number of feature maps. G(y), is the input image, and y = (i, j), represents the position of the pixel value corresponding to the neighbourhood of value x = (i, j), i.e., yϵN(x) in the input image; Here, K_p (m), also with p = 1, 2, ..., 4, belongs to the value in the pth convolution filter in the corresponding position of y, and m = (1, 1), (1, 2)..., (4, 4) means the position of elements in the convolution filter.

C. ReLu Layers

The 4 feature maps generated from convolution 1 are provided as input to the next ReLU layer. This layer activates the non-linear function to each element of the feature maps using Equation (2).

$$ReLU(x) = f\begin{Bmatrix} x, x \geq 0 \\ 0, x < 0 \end{Bmatrix} \tag{2}$$

D. Pooling Layers

In pooling 1, the rectified feature maps are down-sampled to find the local maxima in the neighborhood, using the max-pooling process. The feature maps are down-sampled using Equation (3).

$$D_p(z) = \max\ \{FM_p(x)\}_{\forall x \epsilon N(z)} \tag{3}$$

Here, $D_p(z)$ means the outputs of the pooling processes corresponding to the feature maps, FM_p. The feature map element at x = (i, j) is belonging to the neighborhood of the value of z = (i, j) , i.e., xϵN(z) in the down-sampled feature map. The down-sampled features generated from pooling 1 are provided as input to the next convolution layer 2. The three processes mentioned above such as convolution, ReLU and pooling are repeated in the second and the third layers of CNN. In convolution 2, 6 filters are applied for the convolution process so as to extract feature information from the faces. 24 feature maps are generated from convolution 2 and the features down-sampled in pooling 2.

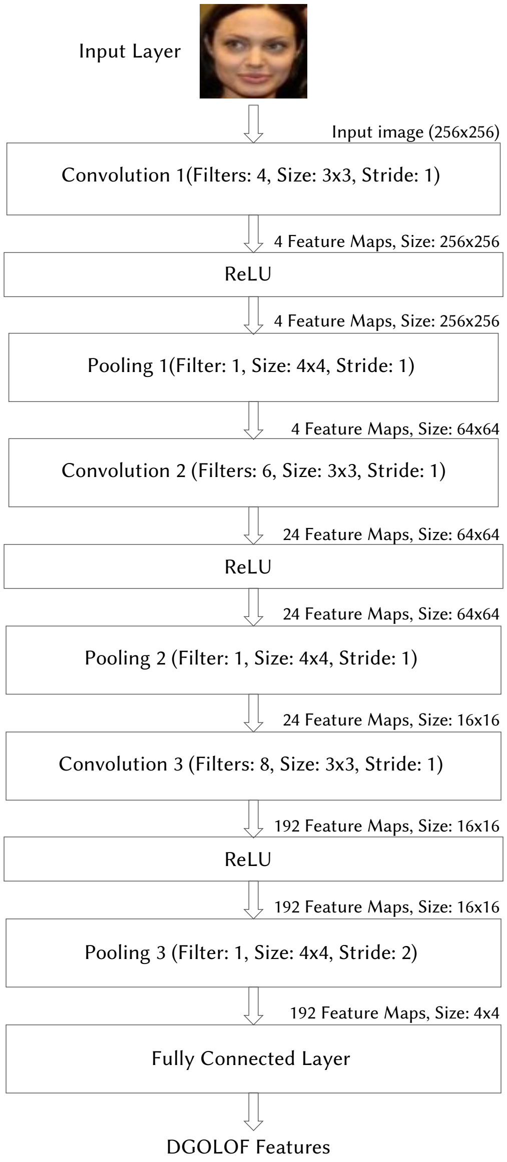

Input Layer
Input image (256x256)
Convolution 1(Filters: 4, Size: 3x3, Stride: 1)
4 Feature Maps, Size: 256x256
ReLU
4 Feature Maps, Size: 256x256
Pooling 1(Filter: 1, Size: 4x4, Stride: 1)
4 Feature Maps, Size: 64x64
Convolution 2 (Filters: 6, Size: 3x3, Stride: 1)
24 Feature Maps, Size: 64x64
ReLU
24 Feature Maps, Size: 64x64
Pooling 2 (Filter: 1, Size: 4x4, Stride: 1)
24 Feature Maps, Size: 16x16
Convolution 3 (Filters: 8, Size: 3x3, Stride: 1)
192 Feature Maps, Size: 16x16
ReLU
192 Feature Maps, Size: 16x16
Pooling 3 (Filter: 1, Size: 4x4, Stride: 2)
192 Feature Maps, Size: 4x4
Fully Connected Layer
DGOLOF Features

Fig. 2. Deep feature extraction using CNN.

A ReLU is comprised between the convolution 2 and the pooling 2 operation. The features of pooling 2 are given as input to the convolution 3. In convolution 3, 8 filters are applied to the convolution process. There are 192 feature maps are generated in convolution 3 and applied to ReLU process. The linearly rectified features are down-sampled in pooling 3. The filter size of 4 × 4 and stride of 1 are applied to all convolution and poling layers. Finally, the deep features are obtained from the fully connected layer of the CNN.

The CNN architecture is trained using a widely used gradient descent method, called stochastic gradient descent (SGD). Table I shows the parameters of the CNN architecture designed in this work for deep feature extraction. By this way, the deep feature of faces is given to the proposed face annotation method.

TABLE I. Parameters for CNN Model

Layer name	No. of filters	Filter size	Stride/ padding	No.of feature maps	Output size
Input layer	n/a	n/a	n/a	1	256x256
Convolution 1	4	3x3	1/0	4	256x256
ReLU	n/a	n/a	n/a	4	256x256
Pooling 1	1	4x4	1	4	64x64
Convolution 2	6	3x3	1/0	24	64x64
ReLU	n/a	n/a	n/a	24	64x64
Pooling 2	1	4x4	1	24	16x16
Convolution 3	8	3x3	1/0	192	16x16
ReLU	n/a	n/a	n/a	192	16x16
Dropout	n/a	n/a	n/a	192	16x16
Pooling 3	1	4x4	2	192	4x4
Fully connected	n/a	n/a	n/a	n/a	3072

The CNN feature enriches spatial localization and effectively exploits minute texture information to resolve real-time issues affecting face images. The convolution and pooling layers of CNN are able to obtain enough information such as edges, orientations, and corner features from the facial images. Edge filters help identify difficult structures caused by facial images. When a face is rotated, key texture features like the eyes, nose and mouth (i.e., non-frontal face) are likely to be lost, but orientation filters help identify enough information from the rest of the face. When elderly faces are considered, corner features help identify the (key point localization) shape of the mouth, nose, eyes and cheeks better than other textures, and effectively differentiate between such faces and other faces. Fig. 3 shows the sample of feature maps generated from the convolutional layers of CNN.

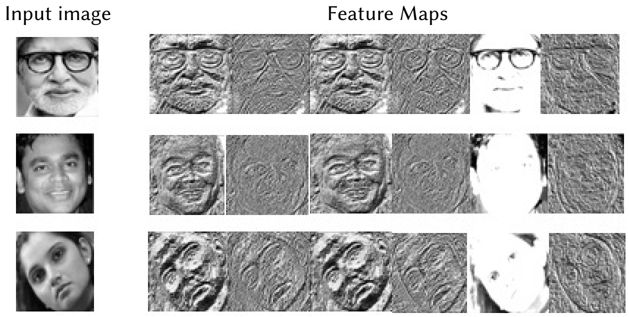


Fig. 3. A sample of CNN feature maps.

III. The Proposed Similarity Matrix-based Noise Label Refinement (SMNLR)

In this section, a new method called the Similarity Matrix-based Noise Label Refinement (SMNLR) is proposed for face annotation. Particularly, two different learning schemes are introduced to obtain two discriminative similarity matrices by learning from noisy labeled faces. The two similarity matrices are further combined to produce a fused similarity matrix, and the noisy labels refined, based on the fused affinity matrix. Section III(A) below introduces a new procedure to generate the Cosine-based similarity matrix. Section III(B) below introduces a new procedure to generate the Mahalanobis-based similarity matrix. Section III(C) describes the fusion of the Cosine and Mahalanobis similarity matrices. Section III(D) introduces a noise label refinement process to refine the noisiness of the labeled faces. The fused matrix effectively discovers the noise labels in labeled facial images. Section III(E) and Section III(F) describe the suggested list generation procedure and face naming procedure respectively.

A. Learning the Cosine-based Similarity

This section explains a new procedure to generate the Cosine-based similarity matrix. The collection of facial features is divided into several subsets, based on their names. The mean feature is calculated, from among the features for each subset, to make a set of effective mean features.The first similarity matrix is calculated between each training facial feature and each subset means feature, based on the Cosine distance.

Each face is characterized as a d-dimensional feature vector using the CNN. For an image x_1 being represented whose CNN feature is defined as f_1, the feature group F is shown as expression (4). The CNN features of each face image, x_p, where p = 1, 2, ..., N in the training dataset, X, can be represented as

$$F = \{f_1, f_2, f_3, \dots \dots f_N\}. \tag{4}$$

Here, N is the total number of features in training set. The CNN features of each face image containing 3072 feature values. Hence the limit for N is specified as 3072 .These training features are grouped into subsets, based on the M names, using Equation (5).

$$S_{ji} = \left[\cup_{M=j}^{N} f_{ji}\right] \tag{5}$$

where $S_{ji} = \{f_{11}, f_{12}, \dots \dots f_{MN}\}$ is a subset and j = 1, 2, ... M is the number of subsets based on person names in the training set and i = 1, 2, ... N represents the number of features in each subset.

The mean feature, MF_j, is calculated for each subset using Equation (6).

$$MF_j = \frac{1}{N} \sum_{i=1}^{N} S_{ji} \tag{6}$$

where number of mean feature, $MF_j = [MF_1, MF_2, \dots MF_M]$, is calculated for all subsets.

The first similarity matrix can be calculated using Equation (7). The Cosine similarity is calculated between each face feature, f_i, in the training set and the mean features of each subset, MF_j.

$$SM1_{ij} = f_i^T . MF_j / (\| f_i \| . \| MF_j \|) \tag{7}$$

where $SM1_{ij}$ represents an element in the ith row and jth column of the cosine similarity matrix, SM1. T is the transpose of the distance value.

B. Learning the Mahalanobis-based Similarity Matrix

This section introduces a new procedure to generate the Mahalanobis-based similarity matrix. Like the first similarity matrix, the mean feature of each subset is calculated, but in contrast, here the mean feature is calculated differently. The collection of facial features is evenly partitioned into several subsets, based on their names. For each subset, the most similar nearest neighbours of each feature among the subset are found using the KNN. The set of minimum distances are calculated in each subset. Finally, the new subset is produced and the mean is calculated. The second similarity matrix is calculated, based on the Mahalanobis distance between each training facial feature and each subset mean feature.

The distance, $d(f_x, f_y)$ between feature f_x and its target neighbours, f_y is calculated using Equation (8).

$$d(f_x, f_y) = \sqrt{\sum_{x=1}^{N} (f_x - f_y)^2} \tag{8}$$

where x = 1, 2, ... N is the number of features in the subset and y = 1, 2, ... T is the number of target nearest neighbours. The set of minimum distances, NF, is formed by using the most similar images with the minimum distance value $d(f_x, f_y)$ using Equation (9).

$$NF=\left\{\min\left(d(f_1,f_y)\right)\cup\min\left(d(f_2,f_y)\right)\ldots\ldots\cup\ \min\left(d(f_x,f_T)\right)\right\} \tag{9}$$

where T represents the number of nearest neighbours. The new subset NF = {f_1, f_2, f_3, … … … f_N} is generated, and the process repeated with all other features in other subsets. The mean of each new subset, NMF_j, is calculated using Equation (10).

$$NMF_j = \frac{1}{N}\sum_{i=1}^{N} f_{ji} \tag{10}$$

where N is the number of features in the new subset, NF. The Mahalanobis distance is calculated between each training set facial feature, f_i and the mean features of each new subset, NMF_j using the following Equation (11).

$$SM2_{ij} = \left((f_i - NMF_j)\right)^T C^{-1}(f_i - NMF_j) \tag{11}$$

where $SM2_{ij}$ represents an element of the ith row and jth column of the second similarity matrix, where N is the number of features in the training dataset, M the total number of subsets, f_i the feature of the ith image in the training dataset, NMF_j the mean feature of the jth subset, C^{-1} the inverse covariance matrix, and T the transpose of the distance value.

C. Learning the Fusion of the Cosine and Mahalanobis-based Similarity Matrices

The first similarity matrix, SM1, is learned from Equation 9 and the second, SM2 , from Equation (12). The two are merged to ake a fused similarity matrix. The fused similarity matrix effectively discovers noise labels, since both matrices contain complementary details of the faces and the discriminative relationship between the faces.

$$FSM_{ij} = \alpha SM1_{ij} + (1-\alpha)SM2_{ij} \tag{12}$$

where FSM_{ij} is the fused similarity matrix, and α the normalization parameter in the range [0, 1] For an enhanced of the label refinement process performance, the normalization parameter value of α is fixed at a range between 0 and 1, respectively, throughout the experiments.

D. Noise Label Refinement Process

The initial noisy name label matrix is refined and reassigned the correct labels, based on the similarity measurement of the fused similarity matrix. The noise labels are replaced with their corresponding subsets, based on the minimum distance between each face and the faces in each subset. Hence, each noise-labeled subset is transformed into a fine-tuned labeled subset, and all faces with their corresponding labels can be relied on for face naming. The noise labels are refined using Equation (13).

$$NL_i=\min\left(FSM_{i,1}, FSM_{i,2}, FSM_{i,3,}, \ldots\ldots\ldots\ldots, FSM_{N,M}\right) \tag{13}$$

where $FSM_{i,1}$ is the similarity value between the ith face and 1st subset.

Fig. 4 shows an example of the label refinement process wherein, for instance, the training features are partitioned into three subsets, based on a person's name. Subset 1, Subset 2, and Subset 3 consist of the sample names P1, P2, and P3 respectively. In each subset, the three different labeled samples are represented by three different shapes, such as a circle, triangle, and square respectively. Subset 1 has three noisy labels. The three samples, which are actually of different persons, are ambiguously labeled P1. This means that the three samples are incorrectly grouped in Subset 1, while the images are grouped on the basis of the name. Similarly, Subset 2 and Subset 3 contain three and two ambiguously labeled samples respectively. The noise labels are rearranged in appropriate subsets using the proposed SMNLR method, which efficiently enhances ambiguously labeled faces with the fused Cosine and Mahalanobis matrices.

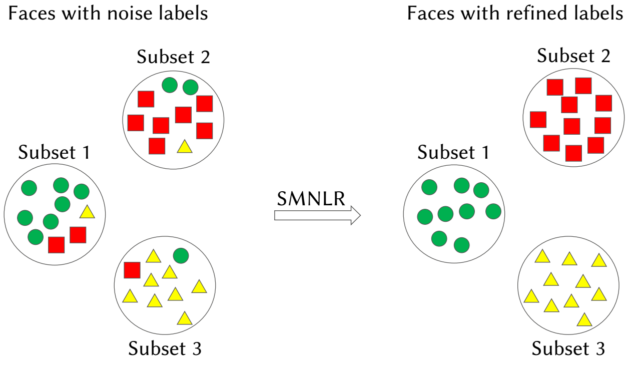


Fig. 4. SMNLR refines the noise labels.

E. Building a Suggested Labels List

In the testing stage, if a corresponding face that is similar to the test face does not occur in the training dataset, it could degrade the face naming capability of the proposed method. To resolve this problem, it is critical to name the unknown face in the test image and, therefore, a suggested label list is created for each instance of the training set. The similarity of each face image and face image collection of all relevant faces are computed. These similarity measurements are sorted in ascending order. The names are retrieved where appears in the labels associated with relevant face images. In training dataset, the suggested labels list, SNL_i is generated for each feature, f_i. The fused similarity matrix, FSM_{ij} is sorted in ascending order using Equation (14).

$$SNL_i = \text{sort}\left(FSM_{i,1}, FSM_{i,2}, FSM_{i,3,}, \ldots\ldots\ldots\ldots, FSM_{i,M}\right) \tag{14}$$

where $FSM_{i,1}$ is the similarity value of the i^{th} training feature corresponding to subset 1, and M is the total number of subsets.

F. Face Naming Using the Multi-class SVM

The face image is annotated with its correct name, using the Multi-class SVM. In the training phase, all the faces with their noise labels are refined, using the proposed method. In the testing phase, the test features are compared with the features of the training set, using the Multi-class SVM classifier. The SMNLR applies the following conditions for face naming:

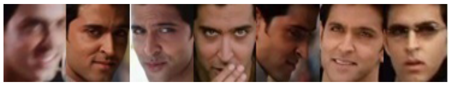
(1) When the multi-class SVM classification result is predicted as positive, a name is assigned to the input face with its corresponding predicted class name.

(2) When the multi-class SVM classification result is predicted as negative, the SMNLR suggests a name list for the input face image.

IV. Database Description

This section describes about the publicly available datasets for face annotation. In this research, the experiments are conducted using the three different datasets, namely, Labeled Faces in the Wild (LFW), Indian Movie Face Database (IMFDB), and Yahoo! News. The LFW dataset is publicly available and it can be collected from http://vis-www.cs.umass.edu/lfw/#explore. IMFDB dataset is publicly available from http://cvit.iiit.ac.in/projects/IMFDB/. Yahoo dataset is available from http://goo.gl/2XlES. It contains the news images with captions. The samples of faces are shown in Table II.

TABLE II. Database Description

Database	Training Set (faces & names)	Testing Set (faces)	Sample face images with various issues
LFW	12500	10450	
Yahoo	8900	7050	
IMFDB	10300	11500	

V. Experimental Results and Analysis

A. The Proposed SMNLR Face Annotation Results for Various Datasets

The performance of the proposed SMNLR method was evaluated with experiments conducted on facial images simulated by noisy labels and real-time challenges. The training set consists of noisy labeled faces, and the testing set of labeled faces. Real-time challenges such as variations in poses, occlusion, illumination and facial expressions are also considered in analysing the effectiveness of the proposed face annotation method. Fig. 5 shows the topmost 5 matching similar faces with their annotation results for 2 sample faces from each dataset.

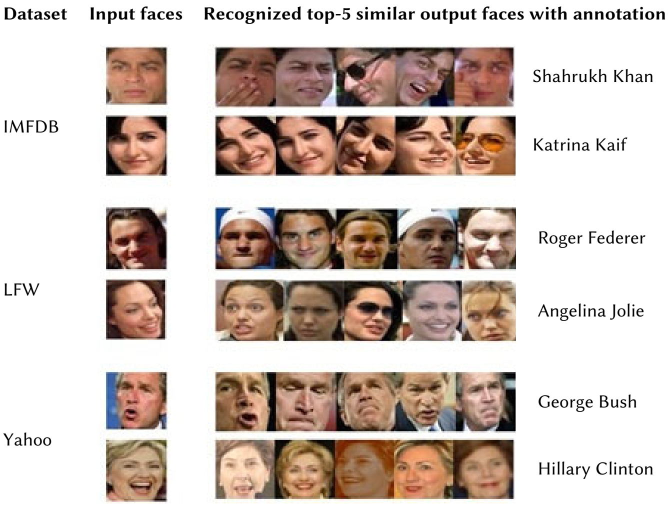


Fig.5. Sample of top-5 recognized similar images with annotation using the proposed method.

VI. Performance Analysis

A. Performance Metrics

The feasibility and effectiveness of the proposed face annotation method is analyzed using the performance metrics given in Equations (15)-(23). The precision, recall and F-score values are calculated using Equations (15), (16) and (17) respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{16}$$

$$\text{F} - \text{score} = 2.\frac{\text{precision}*\text{recall}}{\text{precision+recall}} \tag{17}$$

The recognition rate is validated using Equation (18). The accuracy of the face annotation is evaluated using Equation (19).

$$\text{RecognitionRate} = \frac{\text{Number of correctlymatched images}}{\text{Total test images}} \times 100 \tag{18}$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{19}$$

The true positive rate (TPR) determines the percentage of the face image that is correctly annotated, and is calculated using Equation (20).

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \tag{20}$$

The false positive rate (FPR) typically describes the possibility of falsely naming the input face image, and Equation (21) calculates it.

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{21}$$

The miss rate and error rate of the annotated results are calculated using Equations (22) and (23).

$$\text{Miss rate} = \frac{\text{FN}}{\text{FN+TP}} \tag{22}$$

$$\text{Error rate} = \frac{\text{FP+FN}}{\text{Total}} \tag{23}$$

where TP is true positive, FP is the false positive, TN is the true negative, and FN is the false negative.

B. Fine-tuning the Normalization Parameter, Alpha-(α), for the Proposed SMNLR Face Annotation Method

The noise labels are refined, based on the fused similarity matrix. The fused similarity matrix generation approach uses the normalization parameter, alpha -(α), which is represented in Equation (8). The normalization parameter, α, that combines the two different similarity matrices is experimentally fixed using the three datasets of the LFW, IMFDB, and Yahoo. The impact of the normalization parameter, α, is evaluated in this experiment to find the optimum alpha value. The parameter, α, is set in the range {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}. Table III shows the experimental results.

Certain critical validations are to be drawn from Table III. When setting the value at α=0 and α =1, the performance of the proposed SMNLR method fluctuates, since the noise label refinement procedure becomes ineffective when α=0 and α=1 respectively. This is because, if α is set to 0, the similarity information from the cosine-based matrix can be avoided and the Mahalanobis-based matrix can be quietly updated to handle noise label refinement. At the same time, if α is set to 1, the similarity information from the Mahalanobis-based matrix can be avoided, and the first matrix can be gently updated to carry out noise label refinement. The process of fine-tuning the ambiguity of labeled faces is poorly performed, since the noisy nature of incorrectly labeled faces can be transmitted to correctly labeled faces through the similarity measurements of the fused matrix. Hence, it is clear that the values of 0 and 1 are not applicable to α. After fine-tuning α in

TABLE III. Finding the Optimal Alpha (Α)-value for the Proposed Face Annotation Method

Normalization parameter (α)	Performance of noise label refinement					
	LFW		IMFDB		Yahoo	
	Accuracy (%)	Error rate (%)	Accuracy (%)	Error rate (%)	Accuracy (%)	Error rate (%)
α=0	72	23.8	69	24.7	74	22.6
α=0.1	78	19.5	75	20.7	80	18.1
α=0.2	83	15.3	81	17.5	84	14.7
α=0.3	87	10.4	85	14.4	89	10.6
α=0.4	93	6.8	90	9.3	94	5.2
α=0.5	98	1.3	96	2.2	97	2.1
α=0.6	94	5.6	92	7.1	93	4.7
α=0.7	90	8.3	89	10.5	91	8.3
α=0.8	86	11.2	83	15.8	87	10.7
α=0.9	82	17.5	78	19.4	80	14.5
α=1	74	21.7	70	23.3	78	20.8

TABLE IV. Testing the Performance of the Proposed Method for Varying Proportions of Noisy Labels on Different Datasets

Datasets	Performance Metrics	Proportions of noisy labels (%)									
		10	20	30	40	50	60	70	80	90	100
LFW	Recall	97	97	96	95	94	89	87	85	82	80
	Precision	96	95	95	93	92	90	89	86	85	82
	Accuracy	97	97	96	95	94	91	90	89	87	85
	Miss Rate	2.1	2.7	3.2	4.4	6.8	7.1	7.6	8.1	8.7	10.3
	Error Rate	1.9	2.3	2.9	3.6	5.1	6.4	8.3	9.2	10.5	12.4
IMFDB	Recall	96	94	92	92	90	88	86	83	82	79
	Precision	95	93	93	91	91	89	87	85	83	81
	Accuracy	97	95	92	95	94	93	90	89	86	84
	Miss Rate	2.5	2.9	3.6	4.7	5.9	6.4	8.3	9.6	10.4	11.3
	Error Rate	2.6	3.5	3.8	3.9	5.0	6.9	9.2	10.2	11.8	12.6
Yahoo	Recall	98	97	96	94	91	89	87	86	84	81
	Precision	96	95	93	91	93	92	90	88	85	83
	Accuracy	97	96	95	93	92	90	88	84	82	85
	Miss Rate	1.5	2.7	3.6	5.8	6.5	7.9	8.4	9.7	10.1	12.7
	Error Rate	2.0	2.9	3.1	6.5	7.0	7.3	9.2	10.7	11.8	11.6

the range {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}, it is noted that the proposed method achieves improved results when assigning α to 0.5 on the three datasets, and hence the parameter α is fixed at 0.5. Since the fused similarity matrix comprises several discriminative and prominent details of the cosine and Mahalanobis similarity matrices, it is most effective at exploring noise labels and face naming.

C. Testing the Performance of the Proposed Face Annotation Method Under Different Levels of Noise Labels

Noise label refinement is a difficult issue in face naming. In this experiment, the performance of the proposed face annotation method is tested under different levels of noise labels. Face images and their names are randomly selected from the LFW, IMFDB and Yahoo databases for the training and testing sets. The training dataset contains 12,000 noisy labels of 800 faces. For the purpose of evaluation, different noise levels are simulated in a range from 0% to 100% by updating the randomly-allocated noise labels of each subset in the training set. Here, each level of the noisy labeled faces of all the subsets is applied separately to the proposed method, and the experimental results are shown in Table IV.

Table IV shows that the proposed method refines all the noisy labeled faces perfectly when the noise level ranges from 10% to 30%. When the noise level ranges from 30% to 50%, the proposed SMNLR method reaches 94% accuracy and a lower error rate. When the noise percentage varies from 50% to 100%, it is seen that almost all the noise labels are refined, while still obtaining an accuracy of over 80%. This clearly illustrates the robustness of the proposed SMNLR method in dealing with noisy labeled faces. The SMNLR eliminates the noise labels and re-assigns the correct labels, based on the distance of the least similarity value of each instance. Table IV shows that the SMNLR outperforms different levels of noise, except when the ambiguity percentage is greater than 50%. Hence, the SMNLR achieves enhanced results at low- and middle-levels of noise and becomes vulnerable at high noise levels. The underlying reason for these results is that high ambiguity levels affect the least distance component of the label refining similarity matrix, with the possibility of co-occurrence at such high ambiguity levels.

D. Performance Evaluation of the Proposed Method by Varying Number of Suggested Labels with Respect to the Matching Score

A suggested list is created for each instance of the training set, using the matching score representation. The maximum number of possibilities of extra names for each instance is analysed, based on the matching score. Therefore, this experiment is conducted to find the best combination of matching score levels with size of the suggested labels list. The performance of various combinations of matching score levels with varying sizes of the suggested list is demonstrated in Table V. The matching score levels range from 10% to 50% and the suggested list size that includes 2, 3, 4, 5, 6, 7, 8, 9, and 10 are considered for this experiment, with Table V listing the results.

TABLE V. Performance Evaluation of the Proposed Method By Varying Size of Suggested Labels List and Level of Matching Score

Matching Score level (%)	Datasets	Annotation Accuracy (%)								
		Number of suggestion labels for each face								
		2	3	4	5	6	7	.8	9	10
10	LFW	86.2	90.6	91.7	92.3	93.8	84.7	81.9	80.3	79.2
	IMFDB	85.5	91.2	92.3	93.5	95.9	83.4	82.7	81.6	76.5
	Yahoo	84.4	90.3	91.5	92.1	96.7	86.2	81.5	80.2	77.6
	Average	85.3	90.7	91.8	92.6	95.4	84.7	82	80.7	77.7
20	LFW	90.4	93.4	95.1	95.8	96.6	89.3	86.2	82.3	80.7
	IMFDB	89.1	91.6	94.8	96.9	97.9	86.9	84.5	81.9	79.3
	Yahoo	90.3	90.9	96.5	97.7	97.5	87.6	85.4	83.8	81.5
	Average	89.9	91.9	95.4	96.8	97.3	87.9	85.3	82.6	80.5
30	LFW	84.7	87.2	89.8	94.5	92.8	83.6	78.5	79.3	75.8
	IMFDB	82.4	85.3	90.8	92.2	94.6	81.8	80.2	78.2	76.5
	Yahoo	81.6	84.9	91.4	93.3	93.9	80.5	79.6	77.5	74.2
	Average	82.9	85.8	90.6	93.3	93.7	81.9	79.4	78.3	75.5
40	LFW	79.6	80.2	82.3	84.5	86.8	81.7	76.5	73.3	69.5
	IMFDB	78.3	81.4	82.6	83.1	85.6	80.4	74.7	70.3	70.4
	Yahoo	76.1	80.7	83.1	82.2	83.8	79.3	71.2	72.1	68.6
	Average	78	80.7	82.6	83.2	85.4	80.4	74.1	71.9	69.5
50	LFW	74.7	78.6	80.5	82.4	84.9	72.8	70.4	68.6	66.4
	IMFDB	75.2	77.4	79.3	83.8	85.7	78.5	69.5	67.8	67.2
	Yahoo	74.7	79.2	81.4	81.6	83.9	74.3	71.5	69.3	64.8
	Average	74.8	78.4	80.4	82.6	84.8	75.2	70.4	68.5	66.1

Table V shows that the proposed method produces enhanced results only when the number of suggested labels is less than 6, and decreases progressively as the label size changes from 7 to 10. When the suggested list size ranges from 2 to 6 with a matching score of 10%, the maximum probability of suggested labels for each instance compensates for imbalances in labelling. Consequently, when the suggested list size ranges from 2 to 6 with a matching score of 20%, reliable extra labels for each instance are generated and the annotation performance improved, because of a high probability that. the list of suggested names belongs to the unknown test face. On the contrary, when the number of extra labels ranges from 7 to 10 for each instance, it degrades the performance of the proposed method, and the lower accuracy obtained as a result is noted in Table V. Thus, it is concluded that the number of suggested list sizes is set to a value of 6 with a matching score level of 20%. The suggested list creation procedure enhances the reliability of the SMNLR by building a number of extra labels for each instance, using the fused similarity matrix.

E. Importance of Label Refinement in Face Annotation

In real-life, name labels of faces are incorrect or imperfect, stemming from the manual annotation of online applications. Making face annotation much more reliable by using noise labels is a major issue for real-time face annotation systems. Hence it is essential to refine the label ambiguity of faces without the loss of original labels. To this end, this experiment is conducted to validate the label quality before and after the label refinement process using the TPR, FPR, and accuracy. The values of each are shown in Table VI.

TABLE VI. An Evaluation of the Noise Label Refinement Capability for the Proposed SMNLR Face Annotation Method

Datasets	Without label refinement			With label refinement		
	TPR (%)	FPR (%)	Accuracy (%)	TPR (%)	FPR (%)	Accuracy (%)
LFW	0.60	0.50	0.58	0.97	0.1	0.96
IMFDB	0.52	0.45	0.52	0.96	0.03	0.96
Yahoo News	0.63	0.62	0.54	0.98	0.05	0.97
WDB	0.46	0.55	0.45	0.95	0.1	0.93
Average	0.55	0.53	0.52	0.96	0.07	0.95

The labeling accuracy between noisy labeled faces and refined labeled faces is compared using the TPR, FPR, and accuracy, which reveals contrary results. Table VI proves that the annotated faces with noise label refinement have a high TPR, accuracy value and a low FPR. Further, it clearly reveals that the proposed face annotation method is most reliable and robust.

F. Performance Analysis of the Proposed Face Annotation Method for Different Real-time Challenges

Real-time challenges in face images are commonly a challenge for face annotation. Annotating challenging face images is a difficult task in computer vision, and considerably affects classification and labeling performance. Hence the effectiveness of the proposed SMNLR method is analysed by performing this experiment on expression, occlusion, illumination and pose challenges, using the LFW, IMFDB and Yahoo databases. Table VII displays the performance for SMNLR face annotation against different real life challenging faces.

Table VII clearly shows that the SMNLR method has produced better results for real-time challenges. This is because more than one convolutional filter in the CNN can generate more useful and essential features from the significant facial components such as spatial local contrast, frequency descriptions and orientation properties. In addition to that, the convolution filters use the edges, gradients, directions and corner extraction techniques to obtain more complex features of face image and it overcomes the real-time challenges. However, when compared to normal face recognition, the recognition rate for challenging faces is slightly reduced in terms of expression and occlusion. Since the intrinsic feature information between pixels is not fully extracted from faces, and consequently produces a lower recognition rate.

TABLE VII. A performance Evaluation of the Proposed Face Annotation Method for Real-time Challenges

Datasets	Real-time challenges	Performance Metrics		
		Precision (%)	Recall (%)	Accuracy (%)
LFW	Normal	98.3	96.4	94.4
	Expression	90.5	91.9	86.6
	Illumination	93.6	94.2	90.4
	Occlusion	86.6	84.5	79.7
IMFDB	Normal	97.3	94.4	96.4
	Expression	91.5	90.8	82
	Illumination	95.6	93.3	89.4
	Occlusion	83.7	80.5	79.7
Yahoo	Normal	97.5	96.3	96.8
	Expression	90.5	89.7	81.4
	Illumination	92.7	91.4	88.1
	Occlusion	98.7	97.4	96.8

G. Performance Comparison of the Proposed Face Annotation Method With Existing Methods

To compare and evaluate the effectiveness of the proposed face annotation method with other state-of-the-art-methods, recall and error rate results are displayed in Table VIII. LFW and Yahoo are the most commonly used universal datasets in the face naming community, and are considered for a comparison with all other methods. In all, 4000 samples for training and 3000 samples for testing are taken from LFW, while 5500 samples for training and 4650 samples for testing are taken from the Yahoo dataset. Table VIII shows the experimental results for both the LFW and Yahoo datasets.

TABLE VIII. A comparison of the Proposed Face Annotation Method With State-of-the-art Methods

Face Annotation Methods	LFW		Yahoo	
	Recall (%)	Error rate (%)	Recall (%)	Error rate (%)
Chen's method [3]	78.5	19.4	71.3	21.7
Zeng's method [31]	65.7	23.2	64.1	27.4
Cour's method [32]	74.3	22.5	78.5	23.1
Liu's method [33]	88.1	9.1	90.2	8.4
Su's method [1]	83.4	14.6	80.6	12.9
Kumar's method [35]	90.6	9.2	89.8	10.3
Proposed SMNLR method	97.2	1.9	96.9	2.4

Table VIII clearly demonstrates that the proposed SMNLR method has produced significant results, when compared to state-of-the-art methods. This is because the fused similarity matrix obtains efficient similarity measures between faces with associated noise labels, and eliminates noise labels significantly when resolving the label refinement task. The error rate of the proposed method is also much lower than all other methods. The recall values of the methods of Chen et al. and Cour et al. indicate that their face naming performance is slightly worse than all other methods. The method advanced by Zeng et al. provides a lower recall value and higher error rate because their procedure fails to effectively handle noise labels and other irrelevant information, which impacts annotation results. The methods recommended by Liu et al. and Su et al. achieve recall rates of up to 83.4% and 90.2% respectively. In the methods above, most label ambiguity issues are resolved, and improved results are achieved by comparing them to the methods of Chen et al., Cour et al. and Zeng et al. That's Kumar et al. The method propounded by Kumar et al. produces slightly better results than all other methods, because they employed Convolutional neural networks for feature extraction. Table VIII proves that the proposed SMNLR method outperforms other related state-of-the-art-methods.

VII. Conclusion

In this paper, a new method named as Similarity Matrix based Noise Label Refinement (SMNLR) is proposed for face annotation. Two different similarity matrices can be acquired from first and second similarity matrix learning schemes respectively. In addition, these two matrices are fused to distinguish the uniqueness of faces. Generally, noise labels are refined by using cluster based approaches. On the contrary to existing methods, the proposed SMNLR method effectively exploits the noise label refinement approach for resolving the ambiguity of labels. Since SMNLR is proficient of exploiting the essential minimum distance value representation of faces, it is effective to identify variations within faces. It is noted that the proposed method produced significant results under different level of noisy labeled facial images. It is also observed that the CNNs deep feature offers improved results for annotation. Further, it makes the suggested labels list to overcome the problem of labeling the face that is not occurred in training set. The extensive experiments have been conducted to validate the proposed method using the three databases, such as IMFDB, LFW and Yahoo. The noise labels are synthesized on these three datasets. Moreover, the proposed SMNLR method outperforms various state-of-the-art methods. Finally, it is concluded that the similarity measurements based label refinement approaches can effectively handle the ambiguously labeled facial images for face annotation.

References

[1] X. Su, J. Peng, X. Feng, and J. Wu, "Labeling faces with names based on the name semantic network," *Multimedia Tools and Applications*, vol. 75, no. 11, pp. 6445-6462, 2016.

[2] D. Wang, S. C. Hoi, Y. He, J. Zhu, T. Mei, and J. Luo, "Retrieval-based face annotation by weak label regularized local coordinate coding," *IEEE transactions on pattern analysis and machine intelligence,* vol. 36, no. 3, pp. 550-563, 2014.

[3] C. H. Chen, V. M. Patel, and R. Chellappa, "Learning from ambiguously labeled face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp.1653-1667, 2017.

[4] S. C. Huang, M. K. Jiau, and Y. H. Jian, "Optimisation of automatic face annotation system used within a collaborative framework for online social networks," *IET Computer Vision,* vol. 10, no. 5, pp. 351-360, 2016.

[5] D. Wang, S. S. Hoi, and Y. He, "A unified learning framework for auto face annotation by mining web facial images," *Proceedings of the 21st ACM international conference on Information and knowledge management,* 2012, pp. 1392-1401.

[6] D. Wang, S. C. Hoi, Y. He, and J. Zhu, "Mining weakly labeled web facial images for search-based face annotation," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 1, pp.166-179, 2012.

[7] G. Gao, M. Xu, J. Shen, H. Ma, and S. Yan, "Cast2face: assigning character names onto faces in movie with actor-character correspondence," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 26, no. 12, pp.2299-2312, 2015.

[8] M. Zhu, N. Wang, X. Gao,J. Li J, and Z. Li, "Face Photo-Sketch Synthesis via Knowledge Transfer," *In Proceedings of the Twenty-Eight International Joint Conference on Articial Intelligence (IJCAI),* 2019, pp. 1048-1054.

[9] B. Shikha, P. Gitanjali, and D. Pawan Kumar, "An Extreme Learning Machine-Relevance Feedback Framework for Enhancing the Accuracy of a Hybrid Image Retrieval System," *International Journal of Interactive Multimedia & Artificial Intelligence,* vol. 6, no. 2, pp. 15- 27, 2020, doi: 10.9781/ijimai.2020.01.002.

[10] H. Ito, and H. Koshimizu, "Face image retrieval and annotation based on two latent semantic spaces in fiars," *In Eighth IEEE International Symposium on Multimedia (ISM'06),* 2006, pp. 831-836.

[11] A. Kasthuri, and A. Suruliandi, "A survey on face annotation techniques," *In 4th International Conference on Advanced Computing and Communication Systems (ICACCS),* IEEE, 2017, pp. 1-9.

[12] Y. Yang, Y. Liu, and J. Liu, "Automatic face image annotation based on a single template with constrained warping deformation," IET Computer Vision, vol. 7, no. 1, pp.20-28, 2013.

[13] J. Zhu, S.C. Hoi, and M. R. Lyu, "Face annotation using transductive kernel fisher discriminant," *IEEE Transactions on Multimedia,* vol. 10, no. 1, pp.86-96, 2007.

[14] H. Zitouni, M. F. Bulut, and P. Duygulu, "Recognizing faces in news photographs on the web," *In 2009 24th International Symposium on Computer and Information Sciences, IEEE,* 2009, pp. 50-55.

[15] G. B. De Souza, D. F. da Silva Santos, R.G. Pires, A. N. Marana, and J. P. Papa, "Deep texture features for robust face spoofing detection," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 64, no. 12, pp.1397-1401, 2017.

[16] K. Tang, X. Hou, Z. Shao, and L. Ma, "Deep feature selection and

projection for cross-age face retrieval," *In 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, 2017, pp. 1-7.

[17] A. Kasthuri, A. Suruliandi, and S. P. Raja, "Gabor-oriented local order feature-based deep learning for face annotation," *International Journal of Wavelets, Multiresolution and Information Processing, vol.* 17, no. 05, p. 1950032, 2019, doi.org/10.1142/S0219691319500322.

[18] K. Anburajan, S. Andavar, and P. Elango, "An Empirical Evaluation of Name Semantic Network for Face Annotation," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 13, no. 4, pp.557-571, 2020.

[19] V. Rihani, A. Bhandari, and C. P. Singh, "Face Recognition Using Convolution Filters and Neural Networks," *In IC-AI,* 2006, pp. 185-190.

[20] N. Venkatesh, M. Subhransu, and R. Manmatha, "Automatic image annotation using deep learning representations", *In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM* , 2015.

[21] R. Wang, Y. Xie, J. Yang, L. Xue, M. Hu, and Q. Zhang, "Large scale automatic image annotation based on convolutional neural network," *Journal of Visual Communication and Image Representation*, vol. 49, pp.213-224, 2017.

[22] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," *In 2016 IEEE winter conference on applications of computer vision (WACV)*, 2016, pp. 1-8.

[23] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp.2884-2896, 2018.

[24] M. Khari, A. K. Garg, R. G. Crespo, and E. Verdú, "Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 5, no. 7, p. 22, 2019, doi: 10.9781/ijimai.2019.09.002.

[25] M. S. Maheshan, B. S. Harish, and N. Nagadarshan, "A Convolution Neural Network Engine for Sclera Recognition," *International Journal of Interactive Multimedia & Artificial Intelligence,* vol. 6, no. 1, pp. 78-83, 2020, doi: 10.9781/ijimai.2019.03.006.

[26] L. Celona, S. Bianco, and R. Schettini, "Fine-grained face annotation using deep multi-task CNN," *Sensors,* vol. 18, no. 8, p. 2666, 2018.

[27] W. Jiang, and W. Wang, "Face detection and recognition for home service robots with end-to-end deep neural networks," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2017, pp. 2232-2236

[28] X. Sun, P. Wu, and S.C Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp.42-50, 2018.

[29] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3767-3780, 2019.

[30] O. C. Kurban, T. Yildirim, and A. Bilgiç, "A multi-biometric recognition system based on deep features of face and gesture energy image," In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA,)* 2017, pp. 361-364

[31] Z. Zeng, S. Xiao, K. Jia, T. H. Chan, S. Gao, D. Xu, and Y. Ma, "Learning by associating ambiguously labeled images," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2013, pp. 708-715.

[32] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels", *The Journal of Machine Learning Research*, vol. 12, pp. 1501-1536, 2011.

[33] X. Liu, S. Li, M. Kan, S. Shan, and X. Chen, "Self-error-correcting convolutional neural network for learning with noisy labels," *In 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017),* 2017, pp. 111-117.

[34] X. Su, J. Peng, X. Feng, J. Wu, J. Fan, and L. Cui, "Cross-modality based celebrity face naming for news image collections," *Multimedia Tools and Applications,* vol. 73, no. 3, pp.1643-1661, 2014.

[35] V. Kumar, A. Namboodiri, and C.V Jawahar, "Semi-supervised annotation of faces in image collection," *Signal, Image and Video Processing,* vol. 12, no. 1, pp.141-149, 2018.

[36] V. Maihami, and F. Yaghmaee, "Automatic image annotation using community detection in neighbor images," *Physica A: Statistical Mechanics and its Applications, 507,* 2018, pp.123-132.

[37] M. Zhu, N. Wang, X. Gao, and J. Li, "Deep graphical feature learning for face sketch synthesis," *In Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3574-3580.

[38] M. Zhu, J. Li, N. Wang and X. Gao, "A Deep Collaborative Framework for Face Photo–Sketch Synthesis," In *IEEE Transactions on Neural Networks and Learning Systems* vol. 30, no. 10, pp. 3096-3108, 2019.

[39] X. Wang, S. Wang, J. Wang, H. Shi, and T. Mei, "Co-mining: Deep face recognition with noisy labels," *In Proceedings of the IEEE international conference on computer vision,* 2019, pp. 9358-9367.

[40] S.C. Hoi, D. Wang, I.Y. Cheng, E.W. Lin, J. Zhu, Y. He, and C. Miao, "Fans: face annotation by searching large-scale web facial images," *In Proceedings of the 22nd international conference on World Wide Web, ACM*, 2013, pp. 317-320.

[41] S. C. Huang, M. K. Jiau and C. A. Hsu, "A High-Efficiency and High-Accuracy Fully Automatic Collaborative Face Annotation System for Distributed Online Social Networks," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 24, no. 10, pp. 1800-1813, 2014.

[42] J. Han, J. Hu, and W. Deng, "Constrained Spectral Clustering on Face Annotation System," *In: Chinese Conference on Pattern Recognition,* Springer, Singapore*,* 2016, pp. 3-12.

[43] S. Xiao, D. Xu, and J. Wu, "Automatic face naming by learning discriminative affinity matrices from weakly labeled images," *IEEE transactions on neural networks and learning systems,* vol. 26, no. 10, pp.2440-2452, 2015.

[44] D. Wang, and A. K. Jain, "Face retriever: Pre-filtering the gallery via deep neural net," *In International Conference on Biometrics (ICB), IEEE,* 2015, pp. 473-480.

[45] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," *In Proceedings of the IEEE International Conference on Computer Vision,* 2013, pp. 2408-2415.

[46] H. V. Nguyen, and L. Bai, "Cosine similarity metric learning for face verification," *In: Asian conference on computer vision, Springer, Berlin, Heidelberg*, 2010, pp. 709-720.

[47] Y. Ji, T. Lin, and H. Zha, "Mahalanobis distance based non-negative sparse representation for face recognition," *In International Conference on Machine Learning and Applications, IEEE,* 2009, pp. 41-46.

[48] E. Mostafa, A. M. Ali, and A. A. Farag, "Learning a non-linear combination of Mahalanobis distances using statistical inference for similarity measure," *IET Computer Vision*, vol. 9, no. 4, pp.541-548 2015

[49] M. Wang, Z. Wang, and J. Li, "Deep convolutional neural network applies to face recognition in small and medium databases," *In 4th International Conference on Systems and Informatics (ICSAI), IEEE*, 2017, pp. 1368-1372.

A. Suruliandi

A. Suruliandi completed his B.E. in Electronics & Communication Engineering in the year 1987 from Coimbatore Institute of Technology, Coimbatore. He completed his M.E. in Computer Science & Engineering in the year 2000 from Government College of Engineering, Tirunelveli. He obtained his Ph.D. in the year 2009 from Manonmaniam Sundaranar University, Tirunelveli. He is working as a professor in the Department of Computer Science & Engineering in Manonmaniam Sundaranar University, Tirunelveli. He is having more than 29 years of teaching experience. He published 50 papers in International Journals, 23 in IEEE Xplore publications, 33 in National conferences and 13 in International conferences. His research areas are remote sensing, image processing and pattern recognition.

A. Kasthuri

A. Kasthuri received her M.Sc Degree in Computer Science & Information Technology from Kamaraj University, Tamilnadu in 2012. She received her M.Phil Degree in Computer Science from Manonmaniam Sundaranar University, Tirunelveli in 2015. She is currently pursuing Ph.D Degree in Computer Science & Engineering in Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu. Her Research interest includes Image Processing, Face Recognition, Person re-identification, Pattern Recognition.

S. P. Raja

S. P. Raja was born in Sathankulam, Tuticorin District, Tamilnadu, India. He completed his schooling in Sacred Heart Higher Secondary School, Sathankulam, Tuticorin, Tamilnadu, India. He completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. His area of interest is image processing and cryptography. He is having more than 14 years of teaching experience in engineering colleges. Currently he is working as an Associate Professor in the department of Computer Science and Engineering in Vellore Institute of Technology, Vellore. He published 38 papers in International Journals, 24 in International conferences and 12 in national conferences. He is an Associate Editor of the International Journal of Interactive Multimedia and Artificial Intelligence.

Music Boundary Detection using Convolutional Neural Networks: A Comparative Analysis of Combined Input Features

Carlos Hernandez-Olivan, Jose R. Beltran, David Diaz-Guerra *

Department of Electronic Engineering and Communications, University of Zaragoza, 50018, Zaragoza (Spain)

Received 26 October 2020 | Accepted 9 September 2021 | Published 25 October 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

The analysis of the structure of musical pieces is a task that remains a challenge for Artificial Intelligence, especially in the field of Deep Learning. It requires prior identification of the structural boundaries of the music pieces, whose structural boundary analysis has recently been studied with unsupervised methods and supervised neural networks trained with human annotations. The supervised neural networks that have been used in previous studies are Convolutional Neural Networks (CNN) that use Mel-Scaled Log-magnitude Spectograms features (MLS), Self-Similarity Matrices (SSM) or Self-Similarity Lag Matrices (SSLM) as inputs. In previously published studies, pre-processing is done in different ways using different distance metrics, and different audio features are used for computing the inputs, so a generalised pre-processing method for calculating model inputs is missing. The objective of this work is to establish a general method to pre-process these inputs by comparing the results obtained by taking the inputs calculated from different pooling strategies, distance metrics and audio characteristics, also taking into account the computing time to obtain them. We also establish the most effective combination of inputs to be delivered to the CNN to provide the most efficient way to extract the boundaries of the structure of the music pieces. With an adequate combination of input matrices and pooling strategies, we obtain an accuracy F_1 of 0.411 that outperforms a current work done under the same conditions (same public available dataset for training and testing).

Keywords

Deep Learning, CNNs, Music Analysis, Music Information Retrieval (MIR), Music Structure, Self-Similarity Matrix (SSM).

DOI: 10.9781/ijimai.2021.10.005

I. Introduction

Music Information Retrieval (MIR[1]) is the interdisciplinary science for retrieving information from music. MIR is a field of research that faces different tasks in automatic music analysis, such as pitch tracking, chord estimation, score alignment or music structure detection. One of the most active communities and references in MIR is the Music Information Retrieval Evaluation eXchange (MIREX[2]). This is the community that every year holds the International Society for Music Information Retrieval Conference (ISMIR). Algorithms are submitted to be tested in MIREX's datasets within the different MIR tasks. Most of the previous results analyzed and compared in this work have been presented in different MIREX campaigns.

The automatic structural analysis or Music Structure Analysis (MSA) of music is a very complex challenge that has been studied in recent years [1], but it has not yet been solved with an adequate accuracy that surpasses the analysis performed by musicians or specialists. This kind of analysis is only a part of the musical analysis, which involves musical aspects like harmony, timbre and tempo, and segmentation principles like repetition, homogeneity and novelty [2]. This automatic music analysis can be faced starting from music representations such as the score of the piece, the MIDI file of the piece, or the raw audio file.

In music, *form* refers to the structure of a musical piece, which consists of dividing the musical pieces into small units, starting with the motifs, then the phrases, and finally the sections that express a musical idea. *Boundary detection* is the first step that has to be done in musical form analysis and must be done before the naming of the different segments depending on the similarity between them. This last step is named *Labelling* or *Clustering*. This task, translated to the most common genre in MIREX datasets, the pop genre, would be the detection and extraction of the chorus, verse, or introduction of the corresponding song. Detecting the boundaries of music pieces consists on identifying the transitions where these parts begin and end, a task that professional musicians do almost automatically by listening a piece of music. This detection of the boundaries in a musical piece is based on the *Audio Onset Detection* task, which is the first step for several higher-level music analysis tasks such as beat detection, tempo estimation, and transcription.

[1] https://musicinformationretrieval.com/index.html

[2] https://www.music-ir.org/mirex/wiki/MIREX_HOME

* Corresponding author.

E-mail addresses: carloshero@unizar.es (Carlos Hernandez-Olivan), jrbelbla@unizar.es (Jose R. Beltran), ddga@unizar.es (David Diaz-Guerra).

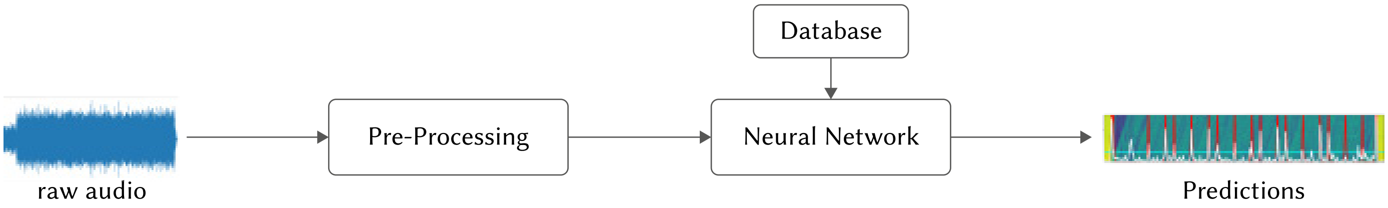


Fig. 1. General scheme of supervised neural networks.

This problem can be accomplished with different techniques that have in common the need of pre-processing the audio files in order to extract the desired audio features and then apply unsupervised or supervised methods. There are several studies where this pre-processing step is made in different ways, so there is not yet a generalized input pre-processing method. The currently *end-to-end* best-performing methods use CNNs trained with human annotations. The inputs to the CNN are MelScaled Log-magnitude Spectograms (MLSs) [3], Self-similarity Lag-Matrices (SSLMs) in combination with the MLSs [4], and also combining these matrices with chromas [5].

One of the limitations of these methods is that the analysis and results obtained depend largely on the database annotator since there can be inconsistencies between different annotators when analyzing the same piece. Therefore, these methods are limited to the quality of the labels given by the annotators and they cannot outperform them.

This paper deals with the issue of structure detection in music pieces. In particular, we study the comparison of different methods of boundary detection between the musical sections by means of Convolutional Neural Networks. The paper is structured as follows: Section II presents an overview of the related work and previous studies in which this work is based on. The Self-Similarity Matrices and the used datasets are also presented. In Section III, the pre-processing method of the matrices that will be used as inputs of the neural network (NN) is explained. Section IV introduces the database used for training, validating and testing, and the labelling process. Section V shows the NN structure and the thresholding and peak-picking strategies and section VI describes the metrics used to test the model and exposes the results of the experiments and their comparison with previous studies. Finally, section VII presents the discussion and section VIII discusses proposals for future lines of work. All code used in this paper, including the pre-trained models of every case of study in this work, is made publicly available[3] and further results are shown in the website[4].

II. Related Work

Several studies have been done in the field of structure recognition in music since Foote introduced the self-similarity matrix (SSM) in 1999 [6] and later, in 2003, he derived from it the selfsimilarity lag matrix (SSLM) [7]. Before the introduction of the SSMs and SSLMs, the studies were based on processing spectrograms [8], but in recent years it has been demonstrated that SSMs and SSLMs calculated from audio features in combination with spectrograms provide better results. We describe some previous works of both unsupervised and supervised methods which belongs to the MIREX's task: Music Structure Segmentation.

A. Unsupervised Methods

The main idea of most of the unsupervised methods is to extract the musical structure of the music pieces but not necessarily the boundaries between the structure sections.

According to Paulus et al. [9], these methods can be summarized in three approaches based on: novelty, homogeneity and repetition. These approaches are computed with unsupervised Machine Learning algorithms such as genetic algorithms (*fitness functions*), Hidden Markov Models (HMM), *K-means*, Linear Discriminant Analysis (NDA), Decision Stump or Checkerboard-like kernels.

The **Novelty-based** approach consists on the detection of the transitions between contrasting parts [1]. This approach is well-performed using checkerboard-like kernel methods which were introduced by Foote in 2000 [10]. These methods have evolved during the years and it has been found that multipletemporal-scale kernels, as those of Kaiser and Peeters in 2013 [11], outperformed the results of previous works by proposing a fusion of the novelty and repetition approaches.

The **Homogeneity-based** approach is based on the identification of sections that are consistent with respect to their musical properties [1]. These methods use Hidden Markov Models, like Logan and Chu [12], Aucouturier and Sandler [13] and Levy and Schandler [14] or combinations of SSMs like Traile and McFee [15], and McFee and Bello [16].

The **Repetition-based** approach refers to finding recurring patterns. These methods apply a clustering algorithm to the SSMs or SSLMs. They are more applicable for labeling the structural parts of music pieces rather than precise segmentation which is required by boundary detection. Lu et al. in 2004 [17], Paulus and Klapuri in 2006 [18], Turnbull et al. [19], McFee and Ellis [20], and McCallum [21] are examples of this method.

To conclude, we can affirm that unsupervised algorithms are very efficient performing the labelling (clustering) part, but not the boundaries detection task, which is better performed by supervised neural networks which came up in 2014 and are described in section B.

B. Supervised Neural Networks

Supervised neural networks learn from input representations given the ground truth, which are the label annotations of the targets (Fig. 1).

Previous studies of boundary detection used Mel-Scaled Log-magnitude Spectograms (MLS) as the inputs of CNNs [3]. This method was based on *Audio Onset Detection* task [22], which consists on finding the starting points of every musically relevant event in an audio signal, specifically the beginning of a music note. This task can be interpreted as a computer vision problem, like edge detection, but applied to spectrograms instead of images with different textures.

Later on, in 2015, Grill and Schlüter improved their previous work by adding SSLMs, which yielded to better results [4], and the addition of SSLMs with different lag factors to the input of the CNN [5], outperforming this method and reaching the best result to date.

In Tables I and II we show a recap of the results of almost all of the previous works that have been done in boundary detection using both unsupervised and supervised neural networks. Results and algorithms nomenclature in Table I have been extracted from MIREX's campaigns of different years. It must be said that the results obtained

[3] https://github.com/carlosholivan/MusicBoundariesCNN

[4] https://carlosholivan.github.io/publications/2021-boundaries/2021-boundaries.html

TABLE I. Results of Boundary Detection of Previous Studies for "Full Structure" and "Segmentation" Tasks. Only the Best-performing Algorithm in Terms of F-measure of Each Year for A 0.5s Time-window Tolerance Is Shown. The F-measure Is Shown for Different Databases (See Sec.D)

Unsupervised Methods								
Year[5]	**Autors [Ref.]**	**Algorithm**	**Input**	**Method**	**F-measure (F_1) for Testing Databases**			
					MIREX09	**RCW-A**	**RCW-B**	**SALAMI**
2009	Paulus & Klapuri [24]	PK	MFCCs, chromas	*Fitness function*	0.27	-	-	-
2010	Mauch et al. [25]	MND1	MFCCs, Discrete Cepstrum	*HMM*	0.325	0.359	-	-
2011	Sargent et al. [26]	SB-VRS1	Chords estimation	*Viterbi*	0.231	0.324	-	-
2012	Kaiser et al. [27]	KSP2	SSM	*Novelty measure*	0.280	0.366	0.289	0.286
2013	McFee & Ellis [20]	MP2	MLS	*Fisher's Linear Discriminant*	0.281	0.355	0.278	0.317
2014	Nieto & Bello [28]	NB1	MFCCs + chromas	*Checkerboard-like kernel*	0.289	0.352	0.269	0.299
2015	Cannam et al. [29]	CC1	Timbre-type histograms	*HMM*	0.197	0.224	0.203	0.213
2016	Nieto [30]	ON2	Constant-Q Transform Spectrogram	*Linear Discriminant Analysis*	0.259	0.381	0.255	0.299
2017	Cannam et al. [29]	CC1	Timbre-type histograms	*HMM*	0.201	0.228	0.192	0.212
Supervised Neural Networks								
2014	Schlüter et al. [31]	SUG1	MLS	*CNN*	0.434	0.546	0.438	0.529
2015	Grill & Schlüter [32]	GS1	MLS + SSLMs	*CNN*	0.523	0.697	0.506	0.541

TABLE II. Results of Previous Works in Boundary Detection Task for 0.5s Time-window Tolerance. It Is Only Showed the Best F-measure Result of Each Reference for Each Database

Unsupervised Methods								
Year	**Autors [Ref.]**	**Input**	**Method**	**Train Set**	**F-measure (F_1) for Testing Databases**			
					MIREX09	**RCW-A**	**RCW-B**	**SALAMI**
2007	Turnbull et al. [19]	MFCCs, chromas, spectrogram	Boosted Decision Stump	-	-	-	0.378	-
2011	Sargent et al. [34]	MFCCs, chromas	Viterbi	-	-	-	0.356	-
Supervised Neural Networks								
2014	Ullrich et. al [22]	MLS	CNN	*Private*	-	-	-	0.465
2015	Grill & Schlüter [4]	MLS + SSLMs	CNN	*Private*	-	-	-	0.523
2015	Grill & Schlüter [5]	MLS + PCPs + SSLMs	CNN	*Private*	-	-	-	0.508
2017	Hadria & Peeters [35]	MLS + SSLMs	CNN	*SALAMI*	-	-	-	0.291

with unsupervised methods on Table I are not as high as the results obtained with supervised neural networks because, as it has been mentioned in section A, the main goal of the unsupervised methods is not the boundary detection (segmentation) itself but the full structure identification (labelling).

C. Self-Similarity Matrices (SSMs)

The Self-Similarity Matrix [2] is a tool not only used in music structure analysis but also in time series analysis tasks. In these matrices, the different parts of the structure of a music piece can be identified as homogeneous regions. This representation of the structural elements of music analysis leads this matrix and its combination with spectrograms to be the input of almost every model described in sections A and B. For this work, this matrix is important because music is in itself *self-similar*, in other words, it is formed by similar time series.

Self-Similarity Matrices have been used under the name of Recurrence Plot for the analysis of dynamic systems [23], but their introduction to the music domain was done by Foote [6] in 1999 and since then, there have appeared different techniques for computing these matrices. The SSM relies on the concept of self-similarity, which is measured by a similarity function that is applied to the audio features representation. The similarity between two feature vectors *yn* and *ym* is a function that can be expressed as Eq. 1 shows. The result is a *N*-square matrix SSM $\in \mathbb{R}^{N\times N}$ being *N* the time dimension:

$$\mathrm{SSM}(n,m) = \delta(y_n, y_m) \tag{1}$$

where $n, m \in [1, ..., N]$.

The similarity function is obtained by the calculation of a distance between the two feature vectors *y* mentioned before. In the literature, this distance is usually calculated as the Euclidean distance δ_{eucl} or the cosine distance δ_{cos}:

$$\delta_{\mathrm{eucl}}(y_n, y_m) = \|y_n - y_m\| \tag{2}$$

$$\delta_{\cos}(y_n, y_m) = 1 - \frac{u.v}{\|y_n\|.\|y_m\|} \tag{3}$$

where *u* and *v* are time series vectors.

Self-Similarity Matrices can be computed from different audio features representations, such as MFCCs or chromas, and they can also be obtained by combining different frame-level audio features [15]. Once the similarity function has been computed for each pair of audio

[5] https://www.music-ir.org/mirex/wiki/<<year>>:MIREX<<year>>_Results - headland "Music Structure Segmentation Results".

feature vectors and the SSM has been calculated, we can filter the SSM by applying thresholding techniques, smoothing or invariance transposition. The SSM can also be obtained with other techniques such as clustering methods as Serra et al. proposed [33], where the SSM is obtained by applying the *k-nn* algorithm.

After Foote in 1999 defined the SSM, in 2003, Goto [7] defined a variant of the SSM which is known as the Self-Similarity Lag Matrix (SSLM). The SSLM is a matrix that represents the similarities between low-level features of one point in time and points in the past, up to a certain *lag time*. This representation makes possible to plot the relations between past events and their repetitions in the future. Some approaches calculate this SSLM after computing the SSM or the recurrence plot as we show in Eq. 4:

$$\mathrm{SSLM}(i,j) = \mathrm{SSM}_{k+1,j} \tag{4}$$

with $i = 1, ..., N$, $j = 1, ..., L$ and $k = i + j - 2modulus(N)$

The dimensions of this matrix are not $N \times N$ as the SSM, but they are $N \times L$, being L the *lag time factor*. That means that the SSLM is a non-square matrix: $\mathrm{SSLM} \in \mathbb{R}^{N \times L}$.

The choice of the type of audio features representation for computing the SSMs or SSLMs, and the choice of using SSMs or SSLMs is one of the most important steps when solving a MIR task and has to be studied depending on the issue we we want to face.

D. Datasets

Previous works had been tested in the annual Music Information Retrieval Evaluation eXchange (MIREX [36]), which is a framework for evaluating music information retrieval algorithms.

The first dataset of the MIREX campaign for the structure segmentation task was the MIREX09 dataset, consisting on a collection of The Beatles' songs plus another smaller dataset[6]. Beatles dataset have 2 annotation versions, one is Paulus Beatles or Beatles-TUT[7] dataset and the second one is the Isophonic Beatles or Beatles-ISO[8] dataset. The second MIREX dataset was MIREX10, formed by the RWC [37] dataset. This dataset has 2 annotation versions; RWC-A[9] of QUAERO project which is the one which corresponds to MIREX10 and RWC-B[10] [38], which is the original annotated version following the annotation guidelines established by Bimbot el al. [39].

A few years later, the MIREX12 dataset provided a greater variety of songs than the MIREX10 [40]. MIREX12 is a dataset formed by the ”Structural Analysis of Large Amounts of Music Information” (SALAMI[11]) dataset which has evolved in its more recent version, the SALAMI 2.0 database. The analysis of MIREX structure segmentation task was published in 2012 [41]. Our work uses the publicly available SALAMI 2.0 dataset.

III. Audio Processing

This work is based on the previous works of Schuler, Grill et al. [3], [4] who propose a pre-proscessing method to obtain the SSLMs from MFCCs features. We will extend these works by calculating the SSLMs from chroma features and applying also the Euclidean distance that has not been considered in preliminary studies, to compute the SSLMs in order to give a comparison and find the best-performing input to the NN model.

A. Mel Spectrogram

The first step of the pre-processing part is to extract the audio features. To do that, we first compute the the Short-Time-FourierTransform (STFT) with a Hanning window of 46ms (2048 samples at 44.1kHz sample rate) and an overlap of 50% as Grill et al. proposed [4]. Then, we obtain a mel-scaled filterbank of 80 triangular filters from 80Hz to 16kHz and we scale logarithmically the amplitude magnitudes to obtain the mel-spectrogram (MLS). We used the *librosa* library [42] to compute the mel-spectrogram. After obtaining the MLS, we apply a max-pooling of $p = 6$ in the temporal dimension to give the Neural Network a manageable size input. The size of the MLS matrix is $P \times N$ with P being the number of frequency bins (that are equal to the number of triangular filters) and N the number of time frames. We define $\mathbf{x}_i$ with $i = 1 \dots N$ as the i-th frame of the MLS.

B. Self-Similarity Lag Matrix From MFCCs

The method that we used to generate the SSLMs[12] is the same method that Grill and Schluter used in [4] and [5], which in turn derives from Serra et al. [43].

The first step after computing each frame mel-spectrogram $\mathbf{x}_i$ is to pad a vector Φ with noise of -70dB with a duration of L frames at the beginning of the mel-spectrogram.

$$\check{\mathbf{x}}_i = \Phi \| \mathbf{x}_i \tag{5}$$

where Φ is a matrix of size $L \times P$ whose elements are equal to -70dB.

Then, a max-pool of a factor of p_1 is done in the time dimension as shown in Eq. 6.

$$\mathbf{x}'_i = \max_{j=1 \dots p_1} (\check{\mathbf{x}}_{(i-1)p_1 + j}) \tag{6}$$

After that, we apply a Discrete Cosine Transform of Type II to each frame omitting the first element.

$$\widetilde{\mathbf{X}}_i = \mathrm{DCT}^{(\mathrm{II})}(\mathbf{x}'_i)_{[2 \dots P]} \tag{7}$$

where P are the number of mel-bands.

Now we stack the time frames by a factor m so we obtain the time series in Eq. 8. The resulting $\hat{\mathbf{X}}_i$ vector has dimensions $[(P-1) \cdot \mathrm{m}] \times [(N+L)/p_1]$ where N is the number of time frames before the max-pooling and L the lag factor in frames.

$$\hat{\mathbf{X}}_i = [\widetilde{\mathbf{X}}_i^{\mathrm{T}} \,\|\, \widetilde{\mathbf{X}}_{i+m}^{\mathrm{T}}]^{\mathrm{T}} \tag{8}$$

The final SSLM matrix is obtained by calculating a distance between the vectors $\hat{\mathbf{X}}_i$. In our work, we use two different distance metrics: the Euclidean distance and the cosine distance. This will allow us to make a comparison between them and conclude which SSLM performs better.

Therefore, the distance between two vectors $\hat{\mathbf{X}}_i$ and $\hat{\mathbf{X}}_{i-l}$ using the distance metric δ is

$$D_{i,l} = \delta(\hat{\mathbf{X}}_i \,, \hat{\mathbf{X}}_{i-l}), \quad l = 1 \dots \left\lfloor \frac{L}{p_1} \right\rfloor \tag{9}$$

where δ is the distance metric as defined in Eqs. 2 and 3.

Then, we compute an equalization factor $\varepsilon_{i,l}$ with a quantile κ of the distances $\delta(\hat{\mathbf{X}}_i \,, \hat{\mathbf{X}}_{i-j})$ for $j = 1 \dots \left\lfloor \frac{L}{p} \right\rfloor$

$$\varepsilon_{i,l} = Q_\kappa \left(D_{i,l}, \cdots, D_{i,\left\lfloor \frac{L}{p} \right\rfloor} \| D_{i-l,1}, \cdots, D_{i-l,\left\lfloor \frac{L}{p} \right\rfloor} \right) \tag{10}$$

We now remove the first L/p lag bins in the time dimension of the distances matrix D and in the equalization factor matrix ε, and we apply Eq. 6 with max-pooling factor p_2. Finally we obtain the SSLM applying Eq. 11.

[6] http://ifs.tuwien.ac.at/mir/audiosegmentation.html

[7] http://www.cs.tut.fi/sgn/arg/paulus/beatles_sections_TUT.zip

[8] http://isophonics.net/content/reference-annotations

[9] http://musicdata.gforge.inria.fr

[10] http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation

[11] https://ddmal.music.mcgill.ca/research/SALAMI/

[12] https://github.com/carlosholivan/SelfSimilarityMatrices

$$R_{i,l} = \sigma\left(1 - \frac{D_{i,l}}{\varepsilon_{i,l}}\right) \tag{11}$$

where $\sigma(x) = \dfrac{1}{1+e^{-x}}$.

Once the SSLM has been obtained, we need to pad some noise to the begin and end of the SSLM because the labels which are used to train our model will be given to the NN as Gaussians (see section IV), so the first and last labels need information in their left and right sides respectively. We add the noise to the begin and end of the SSLM and MLS by padding them with γ = 50 time frames of pink noise at the beginning and end of the MLS matrix. Then we then normalized each frequency band to zero mean and unit variance for MLS and each lag band for the SSLMs. Note also that if there are some time frames that have exactly the same values, the cosine distance would give a NAN (not-a-number) value. We avoid this by converting all this NAN values into zero as the last step of the SSLM computation.

C. Self-Similarity Lag Matrix From Chromas

The process of computing the SSLM from chroma features is similar to the method explained in section B. The difference here is that instead of starting with padding the mel-spectrogram in Eq. 5, we pad the STFT. After applying the max-pooling in Eq. 6, we compute the chroma filters instead of computing the DCT in Eq. 7. The rest of the process is the same as described in section B.

All the values of the parameters used to obtaining the SelfSimilarity Matrices are summarized in Table III. In addition to the Euclidean and cosine metrics, and MFCCs and chromas audio features, we will compare two pooling strategies. The first one is to make a max-pooling of factor $p_1 = 6$ to the STFT (from MLS calculation), and to the Chromas or MFCCs for the SSLMs computation, as it is described in Eq. 6. The other pooling strategy is the one showed in Fig. 2, where we first do a pooling of $p_1 = 2$ and then a pooling of $p_2 = 3$ once the SSLMs are obtained. We denote these pooling variants as 6pool and 2pool3 respectively. The total time for processing all the SSLMs (MFCCs and cosine distance) was a factor or 4 faster for 6pool than 2pool3 because by applying a higher padding factor in Eq. 6 the size of the matrices D and ε is much lower so the calculation of these matrices take more time but it also implies a resolution loss that can affect the accuracy of the model as [4] remarks.

TABLE III. Parameter Final Values

Parameter	Symbol	Value	Units
sampling rate	sr	44100	Hz
window size	w	46	ms
overlap	-	50	%
hop length	h	23	ms
lag	L	14	s
pooling factor 6pool	p	6	-
2pool3	p_1	2	-
	p_2	3	-
stacking parameter	m	2	-
quantile	κ	0.1	-
final padding	γ	50	frames

The general schema of the pre-processing block is depicted in Fig. 2.

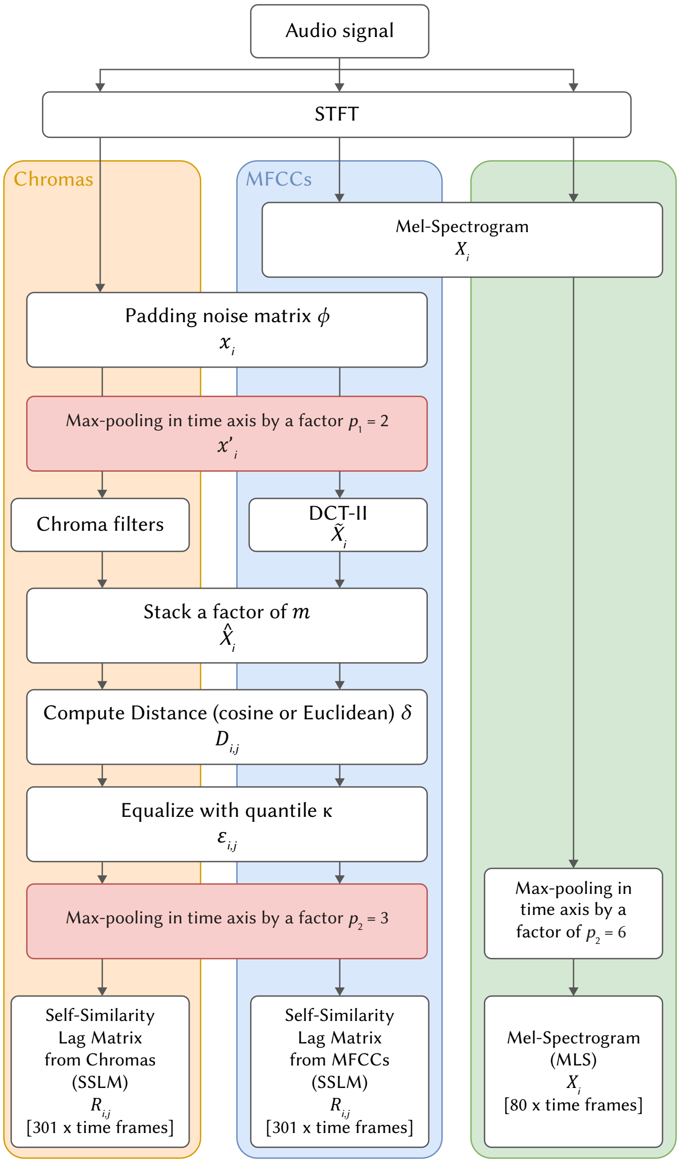

Audio signal
STFT
Chromas
MFCCs
Mel-Spectrogram X_i
Padding noise matrix ϕ x_i
Max-pooling in time axis by a factor p_1 = 2 x'_i
Chroma filters
DCT-II $\tilde{X}_i$
Stack a factor of m $\hat{X}_i$
Compute Distance (cosine or Euclidean) δ $D_{i,j}$
Equalize with quantile κ $\varepsilon_{i,j}$
Max-pooling in time axis by a factor p_2 = 3
Max-pooling in time axis by a factor of p_2 = 6
Self-Similarity Lag Matrix from Chromas (SSLM) $R_{i,j}$ [301 x time frames]
Self-Similarity Lag Matrix from MFCCs (SSLM) $R_{i,j}$ [301 x time frames]
Mel-Spectrogram (MLS) X_i [80 x time frames]

Fig. 2. General block diagram of the pre-processing block in Fig. Each background color contains the steps that are necessary to compute each of the inputs: MLS (green), SSLM from Chromas (orange) and SSLM from MFCCs (blue). The red background in the max-pooling blocks refers to the 2 variants done in this work: 2pool3 is the one showed in the scheme, while 6pool is computed by applying the max-pooling of factor 6 in the first red block and removing the second red block of the scheme.

IV. Dataset

The algorithm was trained, validated and tested on a subset of the Structural Analysis of Large Amounts of Music Information (SALAMI) dataset [44]. SALAMI dataset contains 1048 double annotated pieces from which we could obtain 1006 pieces since the datasest does not provide the audio files due to copyright restrictions. For the training of the model, we used the text files of labels from annotator 1 and for the songs that were not annotated by annotator 1, we use the same text file but from annotator 2.

It is important to highlight that, as described in [35], previous works such as [3], [4] and [5] use a private non-accessible dataset of 733 songs from which 633 pieces were used for training and 100 for validation. Therefore, we re-implemented the work presented in [4] but we trained it in our dataset composed by only public SALAMI pieces and annotations. We split our 1006 SALAMI audio tracks into

65%, 15% and 20%, resulting in 650, 150 and 206 pieces for training, validation and testing respectively.

A. Labelling Process

As explained in [3], it is necessary to transform the labels of the SALAMI text files into Gaussian functions so that the Neural Network can be trained correctly. We first set the center values of the Gaussian functions by transforming the labels in seconds into time frames as showed in Eq. 12 constructing the vector $y1$ which contains the center of the gaussians and has its dimension equal to the number of labels in the text file. In Eq. 12, label$_i$ are the labels in seconds extracted from SALAMI text file "functions" and p_1, p_2, h, sr and γ are defined in Table III.

$$y_i' = \frac{\text{label}_i}{p_1 \cdot p_2} + \frac{h \cdot sr}{\gamma} \tag{12}$$

Then, we apply a gaussian function with standard deviation $\sigma = 0.1$ and μ_i equal to each label value in Eq.12. In Eq.13 we show the expression of the gaussians of the labels.

$$\text{gaussian\_labels}_\text{i} = \mathbf{g}(y_i', \mu_i, \sigma) \tag{13}$$

with

$$\mathbf{g}(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{14}$$

where μ_i is a vector of $\dfrac{y_i \cdot \gamma + \frac{w}{2}}{sr}$ frame from $i = 1 \ldots \left\lfloor \dfrac{N}{p_1 \cdot p_2} \right\rfloor$.

To train the model, we removed the first tag from each text file due to the proximity of the first two tags in almost every file and the uselessness of the Neural Network identifying the beginning of the file. It's also worth mentioning the fact that we have resampled all the songs in the SALAMI database at a single *sampling rate* of 44100Hz as showed in Table III.

V. Model

Our work and current methods that tackle the problem of boundary detection in MSA use neural network-based models that were originally developed for image processing tasks, in particular Convolutional Neural Networks (CNN) [45], [46], [47], [48]. The model developed in this work for boundary detection is shown in Fig. 3. Once the matrices of the pre-processing step are obtained, they are padded and normalized to form the input of a Convolutional Neural Network (CNN). The obtained predictions are post-processed with a peak-picking and threshold algorithm to obtain the final predictions.

A. Convolutional Neural Network

The model proposed in this paper is nearly the same than the model proposed in [3] and [4], so we could compare the results and make a comparison with different input strategies as Cohen [35] did. However, we take into account more inputs combinations and with high and low dimensions in order to see the better inputs combination for the model.

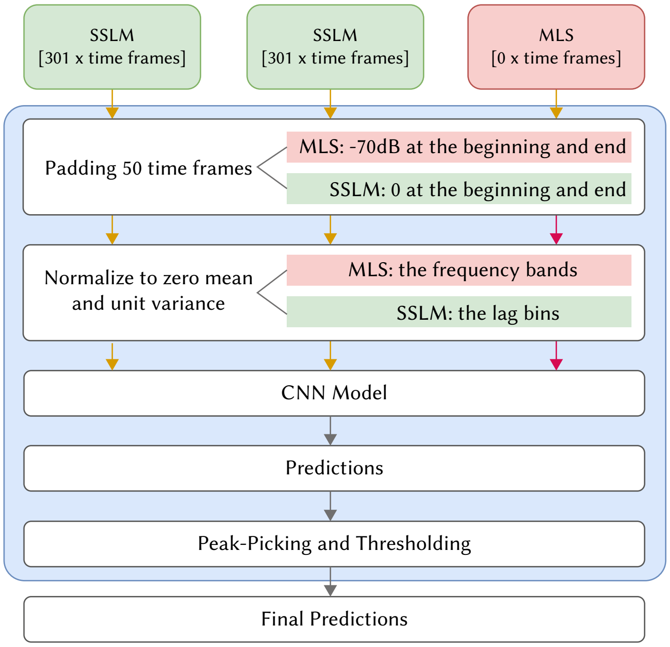


Fig. 3. General block diagram of the Neural Network block in Fig. 1.

The model is composed by a CNN whose relevant parameters are shown in Table IV. The difference between this model and the model proposed in [3] and [4] is that our final two layers are not dense layers but convolutional layers in the time dimension because we do not crop the inputs and get a single probability value at the output, but we give the Neural Network the whole matrix and we obtain a time prediction curve at the output. The general schema of the CNN is shown in Fig. 4.

The parameters of the CNN model have been chosen according to previous literature [4] for a fair comparison in the study of how different input features affect the performance of the MSA task. The changes that have been done from the state-of-the-art model rely on adding the dilation parameter that we use in the layers of our model, and we also changed the last layer of our implementation in comparison with previous literature models. This is because previous studies passed a segment of the SSLM trough the CNN while we pass the entire SSLM to it. The last layer of our implementation outputs one feature map that is passed trough a Sigmoid function which outputs the boundary probability of each time frame of the entire music piece, so the output of the model is a vector of length equal to the time frames of the input. This differs from the literature models where the output is the boundary probability of the segmented part of the input.

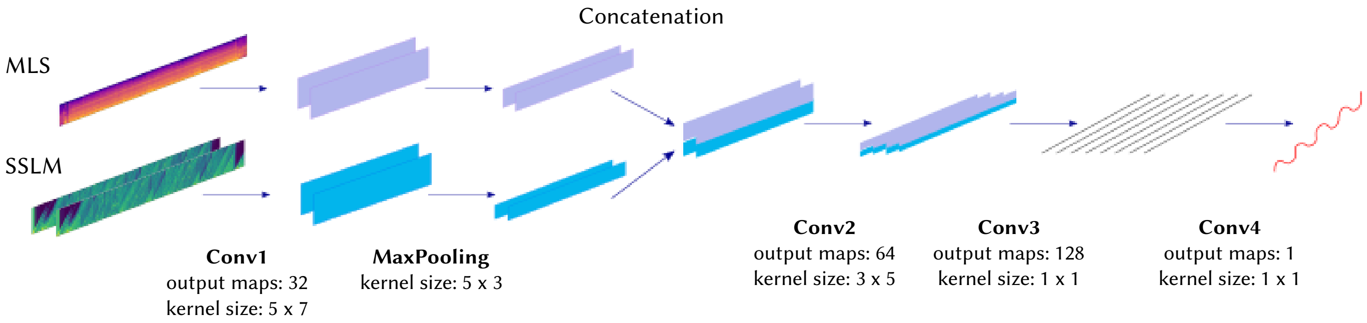


Fig. 4. Schema of the Convolutional Neural Network implemented. The main parameters are presented in Table IV.

TABLE IV. CNN Architecture Parameters of the Schema Presented in Fig. 4

Layer	Parameters
Convolution 1 + Leaky ReLU	output feature maps: 32 kernel size: 5 x 7 stride: 1 x 1 padding: (5-1)/2 x (7-1)/2
Max-Pooling	kernel size: 5 x 3 stride: 5 x 1 padding: 1 x 1
Convolution 2 + Leaky ReLU	output feature maps: 64 kernel size: 3 x 5 stride: 1 x 1 padding: (3-1)/2 x (5-1)*3/2 dilation: 1 x 3
Convolution 3 + Leaky ReLU	output feature maps: 128 kernel size: 1 x 1 stride: 1 x 1 padding: 0 x 0
Convolution 4 + Sigmoid	output feature maps: 1 kernel size: 1 x 1 stride: 1 x 1 padding: 0 x 0

B. Training Parameters

We trained our CNN with *Binary Cross Entropy* or BCEwithLogitsLoss in Pytorch [49] as the loss function which in Pytorch implementation includes a Sigmoid activation function in the last layer of the Neural Network, a *learning rate* of 0.001 and Adam optimizer [50]. We perform *early-stopping* during training to determine the best-performing model. The SSLMs and MLS have to be passed to the GPU one by one because they have different lengths, which means that 1 song is passed forward and backward through the NN at once. However, to get more robust gradients and a more stable optimization process, the optimizer is executed with the average gradients of batchs of 10 songs. We could say that we use a batch size of 1 in terms of GPUs calls but a batch size of 10 in terms of the training. The models were trained on a GTX 980 Ti Nvidia GPU and we used TensorboardX [51] to graph the loss and F-score of training and validation.

C. Peak-Picking

Peak-picking consists on selecting the peaks of the output signal of the CNN that will be identified as boundaries of the different parts of the song. Each boundary on the output signal is considered true when no other boundary is detected within 6 seconds. The application of a threshold helps us to discriminate boundary values that are not higher than an optimum threshold. We calculate the optimum threshold for our experiments by computing the average F_1 in our validation set for all possible threshold values in the range [0, 1] and then we select the highest value. Therefore, the optimum threshold is the value between $[0, 1]$ for which the average F_1 is higher in our validation set. It is reasonable to realise that the optimum threshold value may vary when training our model with the different combination of inputs that we show in Table VI. When we train our model with isolated inputs (see Table V) we compute the threshold with the MLS but we do not vary it when testing SSLMs trainings. We vary the threshold value when we train our model with different inputs combinations in order to optimize the each case of study and give the best-performing method (see Table VI). In Fig. 5, we set a thresold of 0.205 for the models using only the MLS as input and for the rest of the models we used the values indicated in Table VI. From the optimum threshold calculation, we can observe that almost all optimum threshold values for each input variant belong to [2:05; 2:6] Fig. 5 shows Recall, Precision and F-score values (see Section A) of the testing dataset evaluated for each possible threshold value.

TABLE V. Results of Boundaries Estimation According to Different Pooling Strategies, Distances and Audio Features for ± 0:5s and a Threshold of 0.205

Tolerance: ± 0:5s and Threshold: 0.205					
	Input	Epochs	P	R	F1
6pool	MLS	180	0.501	0.359	0.389
	$\text{SSLM}^{\text{MFCCs}}_{\text{euclidean}}$	180	0.472	0.318	0.361
	$\text{SSLM}^{\text{MFCCs}}_{\text{cosine}}$	180	0.477	0.311	0.355
	$\text{SSLM}^{\text{chromas}}_{\text{euclidean}}$	180	0.560	0.228	0.297
	$\text{SSLM}^{\text{chromas}}_{\text{cosine}}$	180	0.508	0.254	0.312
2pool3	$\text{SSLM}^{\text{MFCCs}}_{\text{euclidean}}$	120	0.422	0.369	0.375
	$\text{SSLM}^{\text{MFCCs}}_{\text{cosine}}$	120	0.418	0.354	0.366
Previous works					
2pool3	MLS	-	0.555	0.458	0.465
	$\text{SSLM}^{\text{MFCCs}}_{\text{cosine}}$	-	-	-	0.430

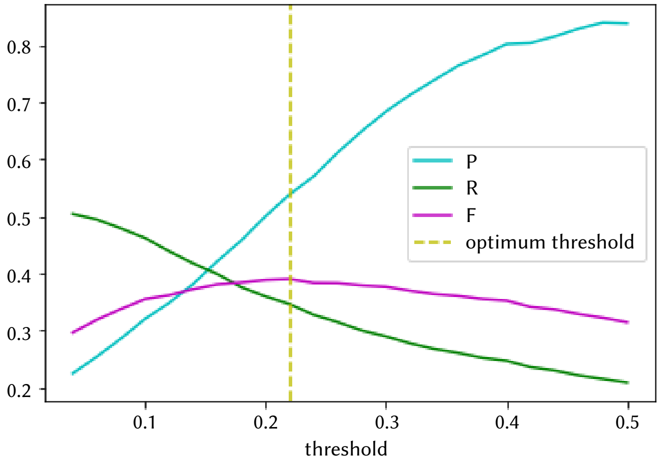


Fig. 5. Threshold calculation through MLS test after 180 epochs of training with MLS.

VI. Experiments and Results

A. Evaluation Metrics

MIREX's campaings use two evaluation measures which are*Median Deviation* and *Hit Rate*. The *Hit Rate* (aslo called F-score or F-measure) is denoted by F_β, where $\beta = 1$ is the measure most frequently used in previous works. Nieto et al. [52] set a value of $\beta = 0.58$, but the truth is that F_1 continues being the most used metric in MIREX works. We will later give our results for both β values. The *Hit Rate* score F_1 is normally evaluated for ± 0:5s and ± 3s time-window tolerances, but in recent works most of the results are given only for ± 0:5s tolerance which is the most restrictive one. We test our model with MIREX algoritm [53] which give us the Precision, Recall and F-measure parameters.

$$\text{Precision: P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{Recall: R} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

TABLE VI. Results of Boundary Estimation With Tolerance ± 0:5s and Optimum Threshold in Terms of F-score, Precision and Recall. Note That Results form Previous Works Did Not Use the Same Threshold Value

Tolerance: ± 0:5s with `2pool3` matrices

Input	Train Database	Epochs	Thresh.	P	R	F_1(std)	$F_{0:58}$
$\mathrm{MLS} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}}$	SALAMI	140	0.24	0.441	0.415	0.402 (0.163)	0.414
$\mathrm{MLS} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}}$	SALAMI	140	0.24	0.428	0.407	0.396 (0.158)	0.404
$\mathrm{MLS} + (\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}} + \mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{euclidean}})$	SALAMI	100	0.24	0.465	0.400	0.407 (0.160)	0.419
$\mathrm{MLS} + (\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}} + \mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{cosine}})$	SALAMI	100	0.24	0.444	0.416	0.404 (0.166)	0.417
$\mathrm{MLS} + (\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}})$	SALAMI	100	0.24	0.445	0.421	0.409 (0.173)	0.416
$\mathrm{MLS} + (\mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{euclidean}} + \mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{cosine}})$	SALAMI	100	0.24	0.457	0.396	0.400 (0.157)	0.420
$\mathrm{MLS} + (\mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{euclidean}} + \mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{cosine}} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}})$	SALAMI	100	0.26	0.526	0.374	**0.411 (0.169)**	**0.451**
End-to-end previous works							
$\mathrm{MLS} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}}$ [4] (2015)	Private	-		0.646	0.484	0.523	0.596
$\mathrm{MLS} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}}$ [35] (2017)	SALAMI	-		0.279	0.300	0.273 (0.132)	-
$\mathrm{MLS} + (\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}} + \mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{cosine}})$ [35] (2017)	SALAMI	-		0.470	0.225	0.291 (0.120)	-

$$\text{F measure: } F_\beta = (1+\beta^2)\frac{P \cdot R}{\beta^2 \cdot P + R} \tag{17}$$

Where:

- TP: True Positives. Estimated events of a given class that start and end at the same temporal positions as reference events of the same class, taking into account a tolerance time-window.
- FP: False Positives. Estimated events of a given class that start and end at temporal positions where no reference events of the same class does, taking into account a tolerance time-window.
- FN: False Negatives. Reference events of a given class that start and end at temporal positions where no estimated events of the same class does, taking into account a tolerance timewindow.

B. Results

1. Isolated Inputs: Distances, Audio Features and Pooling Strategies

We first trained the Neural Network with each input matrix (see Fig. 3) separately in order to know what input performs better. We trained the model using the MLS and SSLMs obtained from MFCCs and Chromas and applying Euclidean and cosine distances, and we also give the results for both of the pooling strategies mentioned before, `6pool` (lower resolution) and `2pool3` (higher resolution). As mentioned in section IV, we removed the first label of the SALAMI text files corresponding to 0.0s label. Results in terms of F score, Precision and Recall are showed in Table V. Note that the results showed from previous works used a different threshold value.

The best-performing input when training our model with isolated inputs is the MLS which has a F_1 value of 0.389 (see Table V). Taking only into account the `6pool` pooling strategy, regarding the SSLMs computed from audio features (MFCCs and chromas) we found that the best-performing SSLMs are the ones that are computed from the MFCCs with more than a 5% difference with the SSLMs computed from chromas.

According to the distance measures with which we compute the SSLMs, we found that there is not a high impact on the results when computing the SSLMs with Euclidean or cosine distances. The F_1 difference between the SSLMs computed with Euclidean or cosine distances is not higher than 1%. Overall, the best-performing SSLM for the `6pool` pooling strategy is the $\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}}$ with a F_1 value of 0.361, which is a 2.8% less than the MLS $F_!$ value of 0.389.

In view of the results in Table V, we can affirm that doing a max-pooling of 2, then computing the SSLMs and doing another max-pooling of 3 afterwards (`2pool3`) slightly improves the results but it does not make a high impact in the performance. The best-performing (`2pool3`) SSLM, the $\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}}$ has a F_1 value of 0.375, which is less than a 2% of the F_1 value of 0.361 for the same SSLM but computed with the `6pool` pooling strategy. This procedure not only takes much more time to compute the SSLMs but also the training takes also much more time and it does not perform better results in terms of F-score.

In Fig. 6 we show an example of the boundaries detection results for some of our input variants on the MLS and SSLMs. We obtained lower results than [4] but higher results than [35] who tried to re-implement [4]. The reasons for this difference could be that the database used by Grill and Schlüter [4] to train their model had 733 non-public pieces. Cohen and Peeters [35], as in our work, trained their model only with pieces from the SALAMI database, so that our results can be compared with theirs, since we trained, validated and tested our Neuronal Network with the same database (although they had 732 SALAMI pieces and we had 1006).

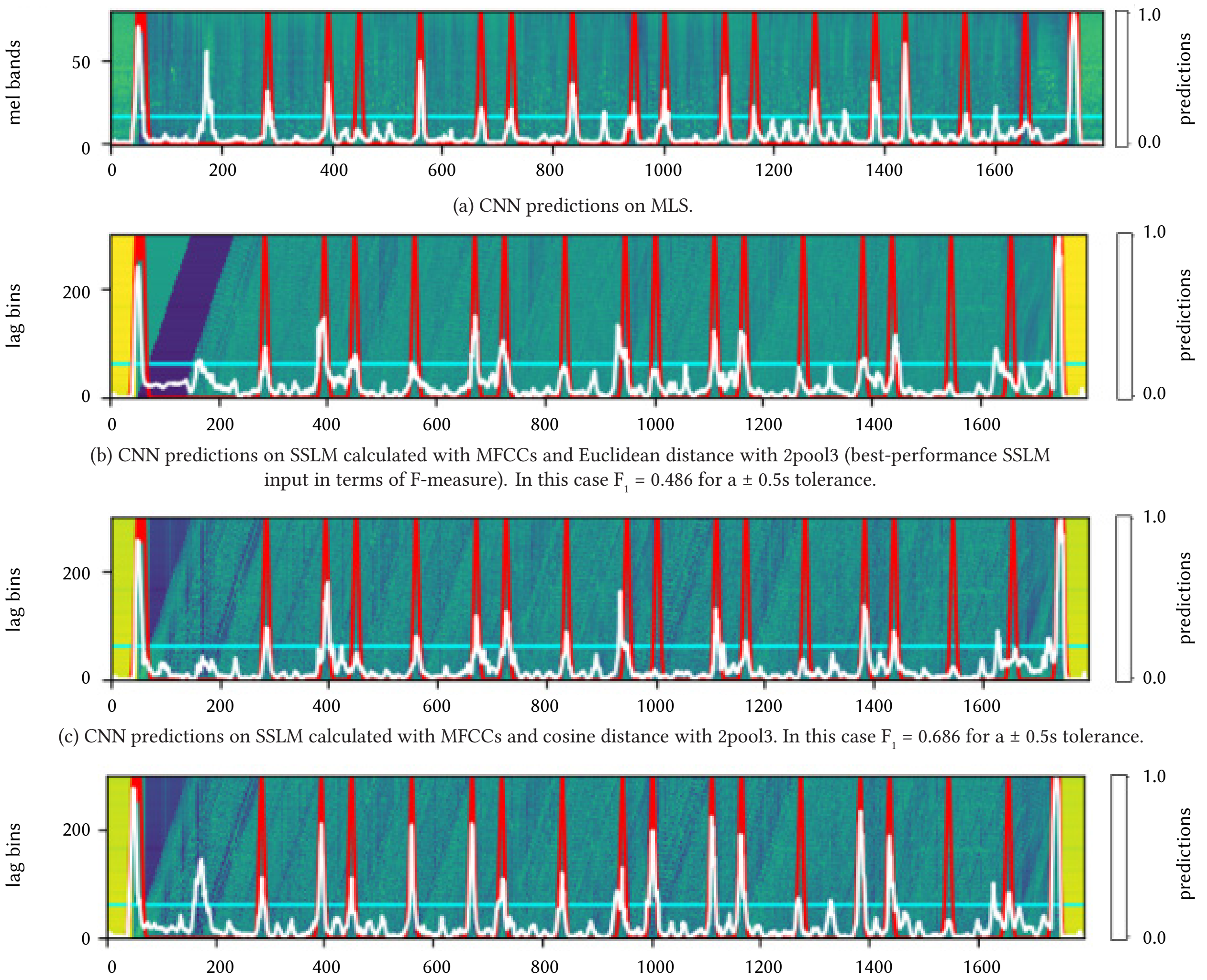
2. Inputs Combination

With the higher results in Table V we make a combination of them as in [4] and later in [35]. A summary of our results can be found in Table VI.

The inputs combination that performs the best in [35] was $\mathrm{MLS} + (\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}} + \mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{cosine}})$ for which $F_1 = 0.291$. We overcome that result for the same combination of inputs obtaining they obtained a F score $F_1 = 0.404$. In spite that, previous works [4] says that cosine distance performs better, we proof that in our model the Euclidean distance gives us better results. We also found that the best-performing inputs combination is $\mathrm{MLS} + (\mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{euclidean}} + \mathrm{SSLM}^{\mathrm{chromas}}_{\mathrm{cosine}} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}} + \mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}})$ for which $F_1 = 0.411$. There is not a huge improvement in the F-measure obtained with this combination in comparison with the results obtained with the combination of the MLS with two SSLMs, but it is still our best result.

VII. Discussion

We can affirm that the best-performing input, when training the model with isolated inputs, is the Mel Spectrogram, which has a F_1 equal to 0.389, more than a 2% higher than the next bestperforming input respresentation, the $\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}}$, whose F_1 is equal to 0.361 (Table V).



(a) CNN predictions on MLS.

(b) CNN predictions on SSLM calculated with MFCCs and Euclidean distance with 2pool3 (best-performance SSLM input in terms of F-measure). In this case F_1 = 0.486 for a ± 0.5s tolerance.

(c) CNN predictions on SSLM calculated with MFCCs and cosine distance with 2pool3. In this case F_1 = 0.686 for a ± 0.5s tolerance.

(d) CNN predictions on SSLM from MFCCs with cosine distance for model MLS + ($\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}}$ + $\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{cosine}}$) . In this case F_1 = 0.75 for a ± 0.5s tolerance.

Fig. 6. Boundaries predictions using CNN on different inputs obtained from the "Live at LaBoca on 2007-09-28" of DayDrug corresponding to the 1358 song of SALAMI 2.0 database. The ground truth from SALAMI annotations are the gaussians in red, the model predictions is the white curve and the threshold is the horizontal yellow line. Note that the prediction have been rescaled in order to plot them on the MLS and SSLMs images. All these images have been padded according to what is explained in the previous paragraphs and then normalized to zero mean and unit variance.

We have also demonstrated that by computing a max-pooling of factor 6 at the beginning of the process not only takes much less preprocessing time but also the training of the Neural Network is faster and it does not affect the results as much as it could be expected. As an example, the $\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}}$ obtained with the `6pool` method has an F_1 value of 0.361 versus the `2pool3` method for the same input which F_1 is equal to 0.375.

Despite the fact that we could not replicate some previous studies of Ullrich et al. [3] and Grill et al. [4] which used nearly the same model that the one which we described in our work, we outperform the results in Cohen et al. [35] work, who also tried to re-implement the model described in the previous literature. There has to be highlighted the fact that previous studies of Ullrich et al. [3] and Grill et al. [4] had at their disposition a private dataset of 733 pieces that they used for training the model, and in this paper the model has been trained only with the public available dataset of SALAMI 2.0.

Adding more inputs to the model does not improve the results in a significant way and it is very time consuming, specially in our last case of study where we take 4 SSLMs in combination with the Mel Spectrogram, which has a $F1$ value of 0.411 in contrast with the $F1$ value of the MLS + $\mathrm{SSLM}^{\mathrm{MFCCs}}_{\mathrm{euclidean}}$ case which is 0.402, so the difference is less than 1%. This leads us to suggest that the use of another neural network architecture that only uses the Mel spectrogram with a SSLM could outperform the current results.

The results obtained in this work improve those presented previously with the same database. However, the accuracy in obtaining the boundaries in musical pieces is relatively low and, to some extent, difficult to use. This makes it necessary, on the one hand, to continue studying different methods that allow a correct structural analysis of music and, on the other hand, to obtain databases that are properly labeled and contain a high number of musical pieces. In any case, the results obtained are promising and allow us to adequately set out the bases for future work.

VIII. Conclusions

In this work we have developed a comparative study to determine the most efficient way to compute the inputs to a convolutional neural network to identify boundaries in musical pieces, combining different methods of generating SSLM matrices. In order to make the comparison and analyse the optimal way to perform the boundary detection task in MSA, different audio features and different pooling strategies have been employed, as well as the combination of different inputs to the CNN.

With an adequate combination of input matrices and pooling strategies, we obtain an accuracy F1 of 0.411 that outperforms the current one obtained under the same conditions (same input data and same datasets for training and testing). In spite of the fact that the best result is given by combining four SSLMs and the MLS, the difference in the F-measure value between our best result and experiments which require less input data and whose training time is lower, is not as high as what it could be expected. We can also affirm that current methods that have been used to date to face music boundary detection do not perform well, so MSA task needs further research because it is not solved yet.

Future work should use new Neural Network architectures that have not been used to solve MSA yet. Architectures employed in language models from Natural Language Processing such as Transformers can lead to out-perform the actual results that are presented in this work due to the memory improvement that they provide in comparison with Long-Short Term Memory Networks (LSTMs). In the case of Transformers, the self-attention mechanism can help the model to better-process the SSMs and SSLMs matrices. Further research, as it has been mentioned before, should also take into account to perform some data augmentation on the current public available datasets in order to have more data to train deep Neural Network models. Data augmentation, if done, should be done with pitch-shifting or by adding Gaussian noise to the inputs, but they should not use rotation or scaling techniques which affect the time distances of the input representations (horizontal axes) and thus, the structure of the music pieces.

References

[1] O. Nieto, G. J. Mysore, C. Wang, J. B. L. Smith, J. Schlüter, T. Grill, B. McFee, "Audio-based music structure analysis: Current trends, open challenges, and applications," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 246–263, 2020.

[2] M. Müller, *Fundamentals of Music Processing - Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[3] K. Ullrich, J. Schlüter, T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, 2014, pp. 417–422.

[4] T. Grill, J. Schlüter, "Music boundary detection using neural networks on spectrograms and self-similarity lag matrices," in *23rd European Signal Processing Conference, EUSIPCO 2015, Nice, France, August 31 - September 4, 2015*, 2015, pp. 1296–1300, IEEE.

[5] T. Grill, J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, 2015, pp. 531–537.

[6] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of the 7th ACM International Conference on Multimedia '99, Orlando, FL, USA, October 30 - November 5, 1999, Part 1*, 1999, pp. 77–80, ACM.

[7] M. Goto, "A chorus-section detecting method for musical audio signals," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, 2003, pp. 437–440, IEEE.

[8] T. Zhang, C. J. Kuo, "Heuristic approach for generic audio data segmentation and annotation," in *Proceedings of the 7th ACM International Conference on Multimedia '99, Orlando, FL, USA, October 30 - November 5, 1999, Part 1*, 1999, pp. 67–76, ACM.

[9] J. Paulus, M. Müller, A. Klapuri, "State of the art report: Audio-based music structure analysis," in *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, 2010, pp. 625–636, International Society for Music Information Retrieval.

[10] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *2000 IEEE International Conference on Multimedia and Expo, ICME 2000, New York, NY, USA, July 30 - August 2, 2000*, 2000, p. 452, IEEE Computer Society.

[11] F. Kaiser, G. Peeters, "Multiple hypotheses at multiple scales for audio novelty computation within music," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 231–235, IEEE.

[12] B. Logan, S. M. Chu, "Music summarization using key phrases," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2000, 5-9 June, 2000, Hilton Hotel and Convention Center, Istanbul, Turkey*, 2000, pp. 749–752, IEEE.

[13] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 360–370, 1999.

[14] M. Levy, M. B. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Speech and Audio Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[15] C. J. Tralie, B. McFee, "Enhanced hierarchical music structure annotations via feature level similarity fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, 2019, pp. 201–205, IEEE.

[16] B. McFee, J. P. Bello, "Structured training for large-vocabulary chord recognition," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, 2017, pp. 188–194.

[17] L. Lu, M. Wang, H. Zhang, "Repeating pattern discovery and structure analysis from acoustic music data," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2004, October 15-16, 2004, New York, NY, USA*, 2004, pp. 275–282, ACM.

[18] J. Paulus, A. Klapuri, "Music structure analysis by finding repeated parts," in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, New York, NY, USA, 2006, p. 59–68, Association for Computing Machinery.

[19] D. Turnbull, G. R. G. Lanckriet, E. Pampalk, M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting," in *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007*, 2007, pp. 51–54, Austrian Computer Society.

[20] B. McFee, D. Ellis, "Dp1, mp1, mp2 entries for mirex 2013 structural segmentation and beat tracking," *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2013.

[21] M. C. McCallum, "Unsupervised learning of deep features for music segmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, 2019, pp. 346–350, IEEE.

[22] J. Schlüter, S. Böck, "Improved musical onset detection with convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 6979–6983, IEEE.

[23] Y. Zou, M. Thiel, M. C. Romano, J. Kurths, "Analytical description of recurrence plots of dynamical systems with nontrivial recurrences," *International Journal of Bifurcation and Chaos*, vol. 17, no. 12, pp. 4273–4283, 2007.

[24] J. Paulus, A. Klapuri, "Music structure analysis with a probabilistic fitness function in MIREX2009," in *Proceedings of the Fifth Annual Music Information Retrieval Evaluation eXchange*, Kobe, Japan, October 2009. Extended abstract.

[25] M. Mauch, K. C. Noland, S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009*, 2009, pp. 231–236, International Society for Music Information Retrieval.

[26] G. Sargent, S. A. Raczynski, F. Bimbot, E. Vincent, S. Sagayama, "A music structure inference algorithm based on symbolic data analysis." MIREX - ISMIR 2011, Oct. 2011. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00618141, Poster.

[27] F. Kaiser, T. Sikora, G. Peeters, "Mirex 2012-music structural segmentation task: Ircamstructure submission," *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2012.

[28] O. Nieto, J. P. Bello, "Mirex 2014 entry: 2d fourier magnitude coefficients,"

Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX), 2014.

[29] C. Cannam, E. Benetos, M. Mauch, M. E. Davies, S. Dixon, C. Landone, K. Noland, D. Stowell, "Mirex 2015: Vamp plugins from the centre for digital music," *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2015.

[30] O. Nieto, "Mirex: Msaf v0. 1.0 submission," 2016.

[31] J. Schlüter, K. Ullrich, T. Grill, "Structural segmentation with convolutional neural networks mirex submission," *Tenth running of the Music Information Retrieval Evaluation eXchange (MIREX 2014)*, 2014.

[32] T. Grill, J. Schlüter, "Structural segmentation with convolutional neural networks mirex submission," *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, p. 3, 2015.

[33] J. Serrà, M. Müller, P. Grosche, J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.

[34] G. Sargent, F. Bimbot, E. Vincent, "A regularity-constrained viterbi algorithm and its application to the structural segmentation of songs," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, 2011, pp. 483–488, University of Miami.

[35] A. Cohen-Hadria, G. Peeters, "Music structure boundaries estimation using multiple self-similarity matrices as input depth of convolutional neural networks," in *AES International Conference Semantic Audio 2017, Erlangen, Germany, June 22-24, 2017*, 2017, Audio Engineering Society.

[36] J. S. Downie, A. F. Ehmann, M. Bay, M. C. Jones, "The music information retrieval evaluation exchange: Some observations and insights," in *Advances in Music Information Retrieval*, vol. 274 of *Studies in Computational Intelligence*, Z. W. Ras, A. Wieczorkowska Eds., Springer, 2010, pp. 93–115.

[37] M. Goto, *et al.*, "Development of the rwc music database," in *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, vol. 1, 2004, pp. 553–556.

[38] M. Goto, "AIST annotation for the RWC music database," in *ISMIR 2006, 7th International Conference on Music Information Retrieval, Victoria, Canada, 8-12 October 2006, Proceedings*, 2006, pp. 359–360.

[39] F. Bimbot, E. Deruty, G. Sargent, E. Vincent, "Methodology and conventions for the latent semiotic annotation of music structure," 2012.

[40] A. F. Ehmann, M. Bay, J. S. Downie, I. Fujinaga, D. D. Roure, "Music structure segmentation algorithm evaluation: Expanding on MIREX 2010 analyses and datasets," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, 2011, pp. 561–566, University of Miami.

[41] J. B. Smith, E. Chew, "A meta-analysis of the mirex structure segmentation task," in *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR, 2013, Curitiba, Brazil*, vol. 16, 2013, pp. 45–47.

[42] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[43] J. Serrà, M. Müller, P. Grosche, J. L. Arcos, "Unsupervised detection of music boundaries by time series structure features," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*, 2012, AAAI Press.

[44] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. D. Roure, J. S. Downie, "Design and creation of a large-scale database of structural annotations," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, 2011, pp. 555–560, University of Miami.

[45] A. A. Alvarez, F. Gómez-Martin, "Motivic pattern classification of music audio signals combining residual and LSTM networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 208–214, 2021.

[46] K. K. Verma, B. M. Singh, H. L. Mandoria, P. Chauhan, "Two-stage human activity recognition using 2d-convnet," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 1–11, 2020.

[47] M. Khari, A. K. Garg, R. G. Crespo, E. Verdú, "Gesture recognition of RGB and RGB-D static images using convolutional neural networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, pp. 22–27, 2019.

[48] S. Jha, A. Dey, R. Kumar, V. K. Solanki, "A novel approach on visual question answering by parameter prediction using faster region based convolutional neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 30–37, 2019.

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett Eds., Curran Associates, Inc., 2019, pp. 8024–8035.

[50] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[51] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: http://tensorflow.org/, Software available from tensorflow.org.

[52] O. Nieto, M. M. Farbood, T. Jehan, J. P. Bello, "Perceptual analysis of the f-measure to evaluate section boundaries in music," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, 2014, pp. 265–270.

[53] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. W. Ellis, "Mir_eval: A transparent implementation of common MIR metrics," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, 2014, pp. 367–372.

Carlos Hernández Oliván

Carlos Hernández Oliván is a PhD student in Electronics at the Universidad de Zaragoza under the supervision of Dr. José R. Beltrán. He received the B.E. and M.Sc. degrees in Industrial Engineering in 2017 and 2019, respectively. He studied viola at the Professional Conservatory of Zaragoza where he received his professional certificate in 2013. He is a researcher at the Department of Electronic Engineering and Communications, University of Zaragoza. His research interests are focused on Music Information Retrieval, in particular, on the music analysis and generation systems with Artificial Intelligence. He is a student member of the International Society of Music Information Retrieval since March 2021.

José R. Beltrán

José R. Beltrán received the M.Sc. and Ph.D. degrees in Physics from the University of Zaragoza, Zaragoza, Spain, in 1988 and 1994, respectively. He is an Associate Professor with the Department of Electronic Engineering and Communications, University of Zaragoza. He has been involved in different research and development projects on Audio Analysis and Processing. His research interests are focused on the study of Automatic Learning Systems for the analysis, processing and synthesis of the musical signal. In 2008, he was a promoter of an academic spin-off: ARSTIC Audiovisual Solutions S.L. devoted to the use of technologies for the artistic and audiovisual fields. Prof. Beltrán is a member of the Aragon Institute for Engineering Research (I3A), Reseach Group in Advanced Interfaces (AffeciveLab).

David Diaz-Guerra

David Diaz-Guerra is a Ph.D. candidate at the University of Zaragoza (Spain), where he received the Bachelor's and Master's degrees in Telecommunications Engineering in 2017 and 2015, respectively. His research focuses on signal processing and machine learning for audio applications.

An Extensive Analysis of Machine Learning Based Boosting Algorithms for Software Maintainability Prediction

Shikha Gupta*, Anuradha Chug

University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, Sector - 16C, Dwarka, New Delhi - 110078 (India)

Received 12 February 2020 | Accepted 16 February 2021 | Published 8 October 2021

Abstract

Software Maintainability is an indispensable factor to acclaim for the quality of particular software. It describes the ease to perform several maintenance activities to make a software adaptable to the modified environment. The availability & growing popularity of a wide range of Machine Learning (ML) algorithms for data analysis further provides the motivation for predicting this maintainability. However, an extensive analysis & comparison of various ML based Boosting Algorithms (BAs) for Software Maintainability Prediction (SMP) has not been made yet. Therefore, the current study analyzes and compares five different BAs, i.e., AdaBoost, GBM, XGB, LightGBM, and CatBoost, for SMP using open-source datasets. Performance of the propounded prediction models has been evaluated using Root Mean Square Error (RMSE), Mean Magnitude of Relative Error (MMRE), Pred(0.25), Pred(0.30), & Pred(0.75) as prediction accuracy measures followed by a non-parametric statistical test and a post hoc analysis to account for the differences in the performances of various BAs. Based on the residual errors obtained, it was observed that GBM is the best performer, followed by LightGBM for RMSE, whereas, in the case of MMRE, XGB performed the best for six out of the seven datasets, i.e., for 85.71% of the total datasets by providing minimum values for MMRE, ranging from 0.90 to 3.82. Further, on applying the statistical test and on performing the post hoc analysis, it was found that significant differences exist in the performance of different BAs and, XGB and CatBoost outperformed all other BAs for MMRE. Lastly, a comparison of BAs with four other ML algorithms has also been made to bring out BAs superiority over other algorithms. This study would open new doors for the software developers for carrying out comparatively more precise predictions well in time and hence reduce the overall maintenance costs.

Keywords

Boosting Algorithms, Feature Selection, Machine Learning, Software Maintainability Prediction, Software Metrics.

DOI: 10.9781/ijimai.2021.10.002

I. Introduction

Software Maintenance, as described in the IEEE Standard for Software Maintenance [1], refers to any modification in a software product after its delivery for improving the performance or any other attribute, for correcting the faults or for adapting the product according to the modified environment. However, software maintenance is not an easy activity because of the complexity that exists in the maintenance behavior of various software systems. Also, a handsome amount of cost is incurred while maintaining software since software maintenance is a high-priced affair. A significant proportion of the comprehensive cost of software during the Software Development Life Cycle (SDLC) is spent in the maintenance phase alone since the cost of maintenance keeps on accumulating with each phase of SDLC. It has been observed that only 30-40% of the resources, including money, time, and effort, are utilized in the development phase, whereas the remaining 60-70% is used for the maintenance activities [1].

* Corresponding author.

E-mail address: shikha.usict.140164@ipu.ac.in

There exists a detailed standard for software quality known as ISO/IEC 25010:2011. It describes eight product quality characteristics, where each characteristic further comprises various sub-related characteristics [2]. Fig. 1 depicts these eight quality characteristics, along with the sub characteristics. Of all the quality characteristics, maintainability is considered for evaluation in the current study since it is one of the most significant characteristics.

In recent times, any software's quality has come out as an essential parameter to account for the software's success. In turn, software quality depends on two main types of attributes: categorized into internal and external categories. Internal attributes like coupling, cohesion, abstraction, inheritance, etc. are directly-measured from the source code during the initial stages of SDLC at the developer level and are hidden from the users. However, external attributes such as durability, understandability, robustness, modifiability, analyzability, etc. are visible to the users and are, in turn, measured indirectly with the help of different internal attributes [3]. The external attributes may also be measured through developers' opinions who write the source code for the open-source software by organizing surveys. However, such surveys involve high costs and are also very time consuming and may produce biased opinions due to the subjectiveness involved in the external quality attributes. Contrarily, measurement of internal quality attributes using Object-Oriented (OO) metric suites has been validated by many researchers for predicting maintainability keeping in view the relationship that exists between the OO metrics & maintainability [4]–[9]. Hence, the current study also uses these OO metrics for Software Maintainability Prediction (SMP).

Software maintainability these days has become one of the essential external attributes of software, which further forms a basis of research for many researchers working in the fields related to software engineering. Software maintainability can be described as the extent to which a particular software system can be changed concerning the number of Lines of Code (LOC). The researchers' fundamental goal is to develop such models for the prediction that are proficient in predicting any software's maintainability accurately and well in advance. This further ensures optimum utilization of resources, including not only money but also the effort and time put in by the development team. Further, the prediction is not the only goal here, but the predictions made should also have high prediction accuracies with the least possible precision errors. Usually, predictions are made with historical data of particular software for which the prediction model is being developed, including both internal and external attributes. A qualitative description of correlation among the internal & external attributes is also found for SMP [3].

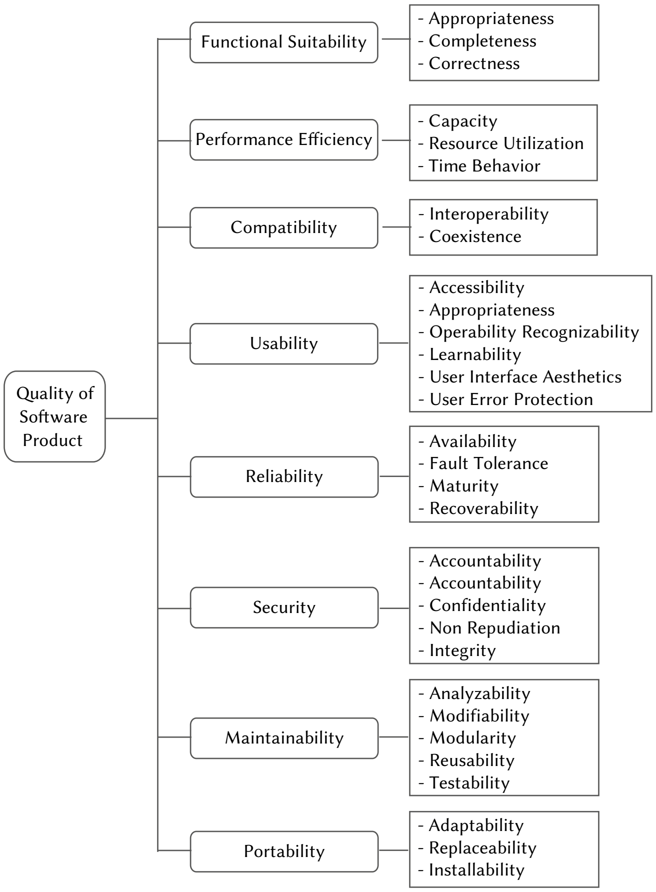


Fig. 1. Model for quality of software product.

In order to estimate the internal quality attributes of software, different metrics have been used such as Afferent Couplings (Ca), Coupling between Object classes (CBO), Efferent Couplings (Ce), & Inheritance Coupling (IC) for coupling; Cohesion among the Methods of a class (CAM), Lack of Cohesion in Methods (LCOM), & Lack of Cohesion among the Methods of a particular Class (LCOM3) for cohesion; Depth of Inheritance Tree (DIT), Number of Children (NOC), & Measure of Functional Abstraction (MFA) for inheritance; Weighted Methods per Class (WMC), & Average Method Complexity (AMC) for complexity; Lines of Code (LOC), & Number of Public Methods (NPM) for size; Response for a Class (RFC) for cardinality; Data Access Metric (DAM) for encapsulation; Measure of Aggregation (MOA) for composition, etc. as compiled in [10], [11]. Software metrics are used to monitor and improve various processes & products in software engineering. These metrics measure various facets of software, such as the design documents or the source code. However, there exist different software metrics based on whether the paradigm is procedural or OO. As per the existing literature, software systems have been analyzed from three perspectives, i.e., the architecture of the system, its design, and the code for SMP [4]. However, out of these, code-level analysis for SMP is the most widely used perspective.

A significant breakthrough for the software industry comes with the advent of Machine Learning (ML). ML [12] is a discipline of artificial intelligence that pertains to the automatic learning capability of different systems and the improvisation of the efficiency based on their past experiences without any explicit programming or learning. The primary focus of ML is developing such programs capable of accessing the data and utilizing it to learn for themselves. The ML process initiates with some data or observations to identify certain patterns in that data, which can further be utilized to make efficient decisions in the future based on the initial data provided. The most important goal of ML is to make the computers capable of automatically learning without human beings' intervention or any kind of external assistance and act accordingly. A wide variety of ML algorithms are currently available for use, broadly classified into two major categories, i.e., supervised (classification & regression) and unsupervised (clustering and association) ML algorithms. Other categories of ML algorithms include semi-supervised and reinforcement learning. Nowadays, ML finds its applications in almost every sphere of life, including web search, computational biology, finance, e-commerce, software engineering, robotics, social networks, debugging, disease diagnosis, stock analysis, marketing analysis, and prediction, etc. Several ML algorithms have been implemented in the fields mentioned above.

As an example, considering the stock analysis field, Sharma et al. [13] in 2018 analyzed ten different supervised classifiers, including logistic regression, C4.5, random forest, etc. for mining of stock data using ICICI bank's data with logistic regression outperforming the other classifiers; Zhong and Enke [14], in 2019, presented a hybrid of deep neural networks with traditional Artificial Neural Networks (ANN) for predicting the return direction for the stock market on a daily basis considering 60 economic & financial features. Rasekhschaffe and Jones [15] in the same year described the primary concepts and the use of ML algorithms in stock selection along with the use of few ensemble models for forecasting stock returns while minimizing the risk of over-fitting. Further, considering the disease diagnosis field, Kaur and Sharma [16], in 2019, extensively reviewed different supervised & nature-inspired ML techniques for mining and analyzing the diagnosis of various psychological disorders using a systematic 3-D search space methodology covering the diagnosis, the disorder & the classification algorithms; and Reddy et al. [17], in 2020, proposed an effective hybrid of adaptive Genetic Algorithm (GA) & fuzzy logic approach (AGAFL) to help the doctors in the early and timely diagnosis of the heart diseases. Again in 2020, Sharma and Kaur [18] conducted a detailed review of the role of several meta-heuristic algorithms based on nature-inspired rules in solving the problem of selecting relevant features for a better classification in different fields; disease diagnosis being the most assessed area. Afterward, in the field of finance, Xiaomeng and Shuliang [19] in 2019, came up with an improved & efficient ML algorithm, i.e., MLIA for the prediction of credit risk in the market of internet finance. Ghoddusi et al. [20] in the same year performed a comprehensive review of various applications of ML in the area of finance or energy economics, including areas like demand forecasting, data processing, risk management, etc., where support vector machines, ANN, and GA stood out to be the most widely used techniques in the concerned field.

Coming back to the field of software maintainability, over the past few years, researchers have developed different ML, evolutionary, & statistical ensemble models for SMP based on the code level metrics intended to predict software maintainability in the most accurate possible manner. Some of the individual ML models developed in the past include General Regression Neural Network (GRNN), Multilayer Perceptron (MLP) [21], Feed Forward 3-Layer Back Propagation Network (FF3LBPN) & Group Method of Data Handling (GMDH) [7]. Models based on nature-inspired algorithms include Evolutionary Algorithms (EA) [22], GA, Functional Link ANN (FLANN), Clonal Selection Algorithm (CSA) & Particle Swarm Optimization (PSO) [23]. Further, the ensemble models developed to date include bagging [8], [24] & boosting, particularly Adaptive Boosting (AdaBoost) [24]. Detailed information about existing models for SMP is provided in the related work section.

Further, ML based boosting techniques have already been explored & implemented successfully in various fields of software engineering comprising software defect prediction [25], software reliability modeling [26] & software fault proneness [27]. However, it is evident from the past studies that none of the researchers has made extensive use of existing Boosting Algorithms (BAs) for SMP apart from one particular BA, i.e., AdaBoost, which still finds a mention in one or two studies [24] for predicting maintainability. Therefore, this study endeavors to predict software maintainability using various BAs and perform an extensive analysis of these BAs for SMP. Boosting is a homogeneous ML ensemble technique proposed by Freund in 1995 [28]. In boosting, an ensemble of classifiers is formed incrementally on adding one classifier at a point in time utilizing the weighted averages so that the base estimators or the weak learners can be converted into strong learners before generating the final output. Unlike other ML algorithms, BAs aim to improvise the prediction capability by training a series of weak learners, each of which compensates for the weaknesses in its preceding learner. BAs are usually intended to reduce the variance and single estimators' bias resulting in a much more stable model. Various BAs are available today that can be used for solving complex and data-driven real-world problems. A significant advantage of using BAs is that these algorithms provide immense powers to the basic ML algorithms, such as Decision Tree (DT), Random Forest (RF), regression, MLP, etc. to improve the prediction accuracy outperforming the basic models. This happens as BAs combine several weak hypotheses of the base estimator that are moderately correct with an objective to derive a notably accurate hypothesis. BAs take several rounds of operation, after which a noteworthy improvement in the accuracy of the training data is achieved. In every round of BAs, samples in the training set are re-weighed, & the base estimator is run on the re-weighted training samples. The BAs main motive is to drive the focus of the weak base estimator towards the error-prone samples. This ultimately leads to the final hypothesis, which is the weak hypotheses' weighted vote. Subsequently, the BAs major strengths include their easy interpretability, availability of feature selection implicitly, resilience towards over-fitting, and strong predictive capability.

Not only this, another significant motivation behind choosing boosting algorithms to conduct this study comes from the effectiveness of the tree boosting that fits the additive tree models having a high ability of representation [29]. This is possible due to the adaptive neighborhood's property, which enables tree boosting use varying degrees for flexibility of different regions in the input space. Hence, it is robust to the dimensionality problem as it performs the feature selection automatically by capturing interactions that are high in order without getting broken. Further, suppose we talk in particular about eXtreme Gradient Boosting (XGB) [29], [30]. In that case, it will learn better structures of trees since these structures determine the neighborhoods & are highly adaptive to the data. XGB uses smart penalization for an individual tree, which can then have a different number of terminal nodes apart from shrinkage. The benefit of using penalization lies in the fact that all the leaves' weight is not shrunk by a common factor. Instead, the weights estimated through fewer pieces of evidence in the data are shrunk even more heavily. Additionally, XGB uses Newton boosting, unlike gradient boosting, & also includes a parameter for randomization to further de-correlate individual trees, which results in the reduction of the overall variance. Also, XGB has a better learning capability. It uses high order approximation at each iteration of the optimization problem & considers the tradeoff between the bias & the variance while fitting the model.

The current study utilizes open-source datasets to analyze various BAs for SMP. The underlying idea behind the selection of open-source datasets for this study was the easy availability of these datasets through various online platforms such as SourceForge & GitHub and the need for generalization and validation of the proposed models for other software in the industry. The study is conducted on a set of seven empirically collected open-source datasets, namely, Abdera, Ivy, jEdit, jTDS, Log4j, Poi, & Rave. A collection of seventeen OO software metrics has been used as the independent variables. In contrast, the maintenance effort (dependent variable) used here is 'Change', which is equal to the number of lines that have been changed per class in the maintenance history. Original datasets are pre-processed to remove those rows where maintenance effort is equal to zero. Feature scaling using MinMaxScaler and feature selection using the Recursive Feature Elimination (RFE) technique is also performed for improving the quality of data. Five different BAs, i.e., AdaBoost, Gradient Boosting (GBM), XGB, LightGBM, and Categorical Boosting (CatBoost), are selected for developing various prediction models for each dataset.

Models are validated using the ten-fold cross-validation technique, and the capability of these models is assessed using Root Mean Square Error (RMSE), Mean Magnitude of Relative Error (MMRE), Pred(0.25), Pred(0.30) & Pred(0.75) taken as the prediction accuracy measures. Friedman test to rank the performance of different BAs used in the study and the Nemenyi test for conducting the post hoc analysis are also performed based on the MMRE. Lastly, a comparison of results achieved on applying the BAs with the results obtained on applying four other ML algorithms (apart from the BAs), viz., DT, MLP, bagging, and Elastic - Net (EN) is also made. Results show that BAs can effectively be applied for SMP, which opens new ways for the researchers to explore these algorithms further. This study's worth lies in the fact that predicting maintainability has become a crucial point of consideration for the software developers throughout the SDLC while developing any software. Also, the software has become a necessity these days since many tasks are becoming automated each day, and this conversion requires some software to be developed. Therefore, the software industry's importance and, in turn, the software is growing leaps and bounds with each passing day. Since the software is being developed, it needs to be maintained also, but as discussed earlier, a handsome amount of cost is required to be spent in the maintenance phase. Thus, some techniques or models are required for predicting software maintainability sufficiently in advance. The current study fulfills this requirement by providing several such models using ML based BAs for a precise prediction of maintainability in good time to help the software developers utilize the resources, such as the money, time & effort judiciously. This would further bring down the maintenance costs associated with any software development process to a considerable extent.

The primary objectives of the current study can be summarized as Research Questions (RQs).

RQ1: Whether BAs can be applied for SMP?

RQ2: Which BA performs the best amongst different BAs based on various prediction accuracy measures for different open-source datasets?

RQ3: What is the comparative performance of various BAs during post hoc analysis when MMRE is taken as an accuracy measure?

RQ4: What is the comparison between the results obtained on applying various ML algorithms (other than the BAs) and the results obtained on implementing BAs?

The remaining paper that follows is organized as - Section II describes the *related work* carried out in the current study field. Section III discusses the datasets, independent variables, & the dependent variables that have been used in the current study. Section IV describes the complete *research methodology*, including pre-processing of datasets, feature scaling & feature selection techniques, description of BAs used, cross-validation technique, prediction accuracy measures, statistical test, and post hoc analysis. *Results and discussions* are described in Section V, whereas Section VI highlights the *threats to validity*. Lastly, Section VII closes the paper with *conclusions & future directions*.

II. Related Work

SMP has become a principal aspect to ascertain any software's quality in the industry over the last few years. Predicting this maintainability in initial stages of development is the need of the hour for efficient and optimum development of any software system. Over the years, substantial research is already being done in the field of SMP by various researchers. They have developed several prediction models using different ML, hybrid, nature-inspired, & other suitable techniques. A summary of these prediction models' details, including the information regarding the types of datasets, metrics suites, ML techniques, validation methods, and the prediction accuracy measures used to compare the performances of the developed prediction models, is provided in Table I. The observed accuracy measures prove that a strong relationship exists between the OO metrics & the maintainability.

Li and Henry [31], in 1993, studied the validation of various OO metrics with the maintenance for the first time using Quality Evaluation System (QUES) & User Interface Management System (UIMS) to prove the existence of a powerful relation among OO metrics and the maintainability. Since then, many researchers have been working in the area of SMP for QUES and UIMS datasets using OO metrics [32]–[38]. Later, Malhotra and Chug [39], in 2012, proposed three ML algorithms, i.e., GMDH, Probabilistic NN (PNN), & GAs, using the Gaussian activation function to predict maintainability & compared their performance with other existing models such as ANN. Results showed that the GMDH model is comparatively more precise & more accurate than the existing models. Again in 2012, Dubey et al. [21] proposed using a robust & adaptive MLP NN model to predict maintainability. MLP, when compared with other models, i.e., WNN & GRNN, was found to be more superior. In another study conducted by Ahmed and Al-Jamini in 2013 [3], fuzzy logic based prediction models, i.e., Mamdani Fuzzy Inference Engine & T-S, were developed & compared for SMP. In comparison, the Mamdani-based prediction model gave the most accurate results of all. In 2014, Malhotra and Chug [7] evaluated the GMDH technique's effectiveness for predicting maintainability by comparing it with the other two techniques, i.e., FF3LBPN & GRNN. It was observed that the GMDH technique performed the best with minimum error & high precision.

In another study, Malhotra and Chug [22] suggested using EAs for SMP using ten-fold cross-validation. The model's performance was analyzed with the help of MRE, MMRE & Pred(q), and later compared with other statistical and ML algorithms. It was found that EAs can effectively predict maintainability with more accuracy and precision as compared to other traditional methods. In 2015, Elish et al. [24] presented three empiric studies for SMP using different homogeneous & heterogeneous ensemble methods. They evaluated and compared three of the heterogeneous ensemble methods to predict maintenance effort, i.e., Weighted-based (WT), Average-based (AVG), and Best in Training-based (BT) ensemble methods. Resultantly, ensemble models came out to be the best when compared to other individual models. All the ensemble and individual models were outperformed by the BT ensemble method. In 2015 only, Kumar et al. [6] suggested using class-level OO software metrics in predicting maintainability with the help of a Neuro-GA for developing the prediction model for QUES and UIMS datasets. Results indicated a successful implementation of Neuro-GA for SMP by generating promising results.

Kumar and Rath [23], in 2016, suggested the use of three Artificial Intelligence (AI) techniques, i.e., FLANN-GA, FLANN-PSO, and FLANN-CSA, to develop models for predicting maintainability along with a few feature reduction techniques. Best & improved results are obtained using feature reduction with FLANN-Genetic. In 2016 again, Chug and Malhotra [9] studied the effect of several ML techniques like GRNN, GMDH, Support Vector Machines (SVM), M5Rules, etc., while predicting the maintainability of seven different open-source software. Results were analyzed for Mean Absolute Error (MAE), RMSE & Pred(q) as the prediction accuracy measures, and it was found that the proposed ML techniques successfully predicted the maintainability for open source software & GMDH and GRNN with Genetic Adaptive Learning (GGAL) performed better than other techniques. In another study conducted by Kumar and Rath [40] in 2017, a Neuro-Fuzzy approach - a hybrid of NN & fuzzy logic was proposed for SMP with Principal Component Analysis (PCA) & Rough Set Analysis (RSA) for selecting suitable features. Results showed that the Neuro-Fuzzy model successfully predicts the software maintainability of OO systems with a further improvement in accuracy using feature selection techniques and parallel computing concepts. In 2018, Baskar and Chandrasekar [41] proved the superiority of the Neuro-PSO (NPSO) model over three other models, namely GMDH, GRNN, & PNN for SMP using MRE, MMRE, & Pred(q) as the accuracy measures. Again in 2018, Alsolai et al. [8] tried to assess the effectiveness of bagging models, i.e., the ensemble models for SMP & proved that there was a noteworthy enhancement in the performance using the bagging models. Further, if combined with *k*-Nearest Neighbour (*k*-NN) as the base model, the bagging model outperformed all the other models resulting in high accuracy.

In 2019, Jha et al. [42] put forth a deep learning approach (LSTM) for SMP using large datasets and 29 OO metrics. They compared the proposed approach with the results of five other ML algorithms, viz. ridge regression, DT, quantile regression forest, SVM, & PCA, to further affirm the LSTM approach's superiority to other models. In the same year, Wang et al. [43] introduced a fuzzy network-based approach for SMP using UIMS & QUES datasets, resulting in an improvement in transparency equal to 71.3% and an improvement in accuracy beyond 11.0%. Recently, in 2020, Gupta and Chug [44] described the cross-project technique for predicting maintainability based on the RMSE values leading to an improvement equal to 13.09% in the overall performance of the predictive models. Again, Gupta and Chug [45] in the same year presented an enhanced RF approach to predict

TABLE I. Summarized Details of Different Prediction Models Developed by Researchers for SMP

Study	Dataset	Metric Suite	Prediction Model	Validation Method	Prediction Accuracy Measure/s (PAMs)	Values for PAMs
Li & Henry [31]	UIMS & QUES	C&K metric suite	MLR	-	-	
Dagnipar & Jahnke [32]	Fujaba-UML (FUML) & Dynamic Object Browser (dobs)	Size-NIM & TNOS, Inheritance-NOC & AID, Cohesion-LCC	Regression Model	LOO	R-square adjusted (between 61.60% & 99.70%)	Between 61.60 & 99.70 %
Thwin & Quah [33]	QUES	DIT, MPC, RFC, LCOM, DAC, WMC, NOM, SIZE1 & SIZE2.	WNN, GRNN	10-cross-validation	R squared, Correlation coefficient r	R^2=WNN-0.56067 & GRNN-0.71139, r= WNN-0.7609805 & GRNN-0.8580623
Koten & Gray [34]	UIMS & QUES	DIT, NOC, MPC, RFC, LCOM, DAC, WMC, NOM, SIZE1, SIZE2	Linear Regression (LR), BNM	10-cross validation	Absolute Residual (Ab.Res.), MRE, MMRE, Pred(q)	Using BNM, for UIMS, MMRE = 0.972, pred(0.25) = 0.446, pred(0.30) = 0.469 For QUES, MMRE = 0.452, pred(0.25) = 0.391, pred(0.30) = 0.430
Aggarwal et al. [35]	UIMS & QUES	LCOM, NOC, DIT, WMC, RFC, DAC, MPC, NOM	ANN	-	MARE, MRE	MARE=0.265, MRE=0.09
Zhou & Leung [36]	UIMS & QUES	WMC, DIT, RFC, NOC, LCOM, MPC, DAC, NOM & SIZE2 & SIZE1	MLR, ANN, RT, SVR, MARS	LOO cross-validation	Residual (Res.), Absolute Residual Error (ARE), MRE, MMRE, Pred(q)	Using MARS, for UIMS, MMRE=1.86, pred(0.25)=0.28, pred(0.30)=0.28, For QUES, MMRE=0.32, pred(0.25)=0.48, pred(0.30)=0.59
Elish & Elish [37]	UIMS & QUES	C&K - WMC, DIT, NOC, RFC, & LCOM; Li & Henry - MPC, DAC, NOM, & SIZE2; & SIZE1	TreeNet classifier	LOO cross-validation	MMRE, MRE, Pred(q), underestimation, overestimation	For UIMS, MMRE=1.57, pred(0.25)=0.31, pred(0.30)=0.41, For QUES, MMRE=0.42, pred(0.25)=0.58, pred(0.30)=0.65
Kaur et al. [38]	UIMS & QUES	LCOM, DIT, WMC, NOC, RFC, DAC, MPC, NOM	ANN, FIS, ANFIS	-	MARE, MRE, R-value, p-value	MARE=36.8% (feed forward ANN), 25.5% (GRNN), 30.8% (FIS), 24.2% (ANFIS)
Malhotra & Chug [39]	UIMS & QUES	WMC, DIT, NOC, RFC, LCOM, MPC, DAC, NOM, Size1, Size2	GMDH, GA, PNN	Hold-out	MRE, MMRE, Pred(q), R-Square, p-value	For GMDH, MMRE=0.210, pred(0.25)=0.69, pred(0.30)=0.722, pred(0.75)=0.944, For GA, MMRE=0.220, pred(0.25)=0.66, pred(0.30)=0.722, pred(0.75)=0.972, For PNN, MMRE=0.230, pred(0.25)=0.68, pred(0.30)=0.75, pred(0.75)=0.944,
Dubey et al. [21]	UIMS & QUES	DIT, NOC, RFC, WMC, LCOM, MPC, DAC, NOM, Size1, Size2	MLP NN	-	R-square, r, MAE, min Absolute Error (AE), max AE	Using MLP. for UIMS, R^2=0.8274, r=0.946, MAE=17.86, for QUES, R^2=0.988, r=0.976, MAE=5.264
Ahmed & Al-Jamini [3]	UIMS & QUES	DIT, NOC, MPC, RFC, LCOM, DAC, WMC, NOM, SIZE1, SIZE2	Fuzzy logic-based models - Mamdani Fuzzy Inference Engine & T-S	-	MRE, Normalized Root Mean square Error (NRMSE), MMRE, Pred(q)	For UIMS, MMRE=0.53, NRMSE=0.21, pred(0.25)=0.30, pred(0.30)=0.35, for QUES, MMRE=0.27, NRMSE=0.16, pred(0.25)=0.52, pred(0.30)=0.62
Malhotra & Chug [7]	FLMS & EASY	WMC, DIT, NOC, RFC, LCOM, MPC, DAC, NOM, SIZE1, SIZE2	GRNN, FF3LBPNN & GMDH	Hold-out	MRE, MMRE, Pred(q), Overestimate, Underestimate	For GRNN, MARE=0.5476, pred(0.25)=0.44, pred(0.30)=0.47, for FF3LBPNN, MARE=0.4578, pred(0.25)=0.51, pred(0.30)=0.59, for GMDH, MARE=0.3566, pred(0.25)=0.61, pred(0.30)=0.71
Malhotra & Chug [22]	Apache Poi & Rave	WMC, DIT, NOC, CBO, RFC, LCOM, LOC	A set of 14 statistical regression, traditional ML & hybrid algorithms	10-fold cross-validation	MRE, MMRE, Pred(0.25), Pred(0.30)	EAs achieved accuracy in the range of 22-25%

Study	Dataset	Metric Suite	Prediction Model	Validation Method	Prediction Accuracy Measure/s (PAMs)	Values for PAMs
Elish et al. [24]	UIMS & QUES for Regression Problem	WMC, DIT, NOC, RFC, LCOM, MPC, DAC, NOM, SIZE1, SIZE2	Different homogeneous & heterogeneous ensemble methods (AVG, WT & BT ensemble)	Ten-fold cross-validation	MMRE, Standard Deviation Magnitude of Relative Error (StdMRE), Pred(q)	Using BT, for UIMS, MMRE=0.97, StdMRE=1.61, pred(0.3)=25, for QUES, MMRE=0.41, StdMRE=0.32, pred(0.3)=60
Kumar et al. [6]	QUES & UIMS	WMC, NOC, DIT, RFC, LCOM, MPC, DAC, NOM, SIZE1, SIZE2	Neuro-GA	10-fold (QUES) and 5-fold (UIMS) cross-validation	MAE, MARE, RMSE, Standard Error of the Mean (SEM)	MMRE=0.3155 (UIMS), 0.3775 (QUES)
Kumar & Rath [23]	QUES & UIMS	WMC, DIT, NOC, LCOM, RFC, MPC, DAC, NOM, SIZE1, SIZE2	FLANN-GA, FLANN-PSO, FLANN-CSA	QUES-10-fold cross-validation, UIMS-5-fold cross-validation	MAE, MMRE, SEM, True Error (e), Estimate of True Error ($\hat{e}$)	Using FGA, MMRE=0.2881 (UIMS), 0.3889 (QUES), using FPSO, MMRE=0.3238 (UIMS), 0.3650 (QUES), using FCSA, MMRE=0.2843 (UIMS), 0.4469 (QUES)
Chug & Malhotra [9]	7 Open Source Software (Drumkit, OpenCV, Abdera, Ivy, Log4j, jEdit, JUnit)	WMC, DIT, NOC, RFC, DAM, MOA, MFA, CAM, AMC, CBO, LCOM, LCOM3, NPM, Ca, Ce, IC, LOC	Thirteen different ML classifiers like LR, M5Rules, GMDH, GRNN, SVM, PNN, etc.	Ten-fold cross-validation	MAE, RMSE, Pred(q)	Pred(0.25) > 60% in all cases using different ML techniques, GGAL & GMDH superior of all techniques
Kumar & Rath [40]	UIMS & QUES	DIT, WMC,RFC, DAC,LCOM, NOC,MPC, NOM,SIZE1, SIZE2	Neuro-Fuzzy Approach & Parallel Computing concept	Five-fold cross validation	MAE, MARE, MMRE, SEM, True Error (e), Estimate of True Error ($\hat{e}$)	MMRE=0.2826 (UIMS), 0.3375 (QUES)
Baskar & Chandrasekar [41]	QUES & UIMS	DIT, WMC, NOC, CBO, LCOM, MPC, RFC, DAC, NOM, Size1, Size2	NPSO	-	MRE, MMRE, Prediction (Pred(q))	MaxMRE=2.02547, MMRE=0.2931, pred(0.25)=0.2998, pred(0.75)=0.5612
Alsolai [8]	QUES	WMC, DIT, NOC, RFC, LCOM, MPC, DAC, NOM, SIZE2, SIZE1	Individual Models (RT, MLP, *k*-NN, M5Rules) & a bagging ensemble model	10-fold cross-validation	MRE, MMRE, Pred(0.25), Pred(0.30), Standard Deviation of Absolute Residuals (SD. Ab.Res.)	Using bagging ensemble models, for RT, MMRE = 0.3, pred(0.25)=0.6, pred(0.30)=0.7, for MLP, MMRE=0.2, pred(0.25)=0.7, pred(0.30)=0.8, for *k*-NN, MMRE=0.1, pred(0.25)=0.9, pred(0.30)=0.9, for M5Rules, MMRE= 0.3, pred(0.25)=0.5, pred(0.30)=0.6
Wang et al. [43]	UIMS & QUES	DIT, NOC, MPC, RFC, LCOM, DAC, WMC, NOM, SIZE2, SIZE1	Fuzzy network	-	MMRE, Transparency (TI)	Best MMRE=0.443, best TI=1
Gupta & Chug [44]	QUES & UIMS	DAC, DIT, LCOM, MPC, NOC, NOM, RFC, SIZE1, SIZE2, WMC	The Cross-Project technique using 19 different regression models	10-fold cross-validation	RMSE	Without CPSMP, Average RMSE=82.31, with CPSMP, Average RMSE=71.53
Gupta & Chug [45]	QUES & UIMS	DAC, DIT, LCOM, MPC, NOC, NOM, RFC, SIZE1, SIZE2, WMC	RF with three different feature selection techniques	10-fold cross-validation	R^2	For QUES, R^2=0.9207; for UIMS, R^2=0.9907

maintainability by combining RF algorithm with three different feature selection methods, i.e., chi-squared, RF, & linear correlation using R^2 as the accuracy estimator. Results show a remarkable improvement in the R^2 values using the enhanced RF approach compared to the basic existent RF approach. Further, Gupta and Chug [46] also propounded an effective utilization of Least Squares SVM (LS-SVM) in predicting maintainability by deriving notable values of MAE, RMSE, & MMRE on using LS-SVM.

It is observable from Table I and from the discussion of various studies conducted in the field of SMP that many researchers have already proposed a vast number of prediction models for SMP to date. However, most of them have used only the publicly available traditional Li and Henry datasets [31] for conducting their research rather than use open-source datasets. It was found that only two of the studies have used open-source datasets [9], [22]. Also, none of the researchers has made extensive use of ensemble methods, particularly the boosting techniques with any kind of dataset, to predict software maintainability apart from a few who considered ensemble models for SMP in their study [8], [24]. Hence, to overcome the above-identified gaps of the existing studies and due to the motivation gained through the availability and effectiveness of various BAs as discussed in the Introduction, the current study attempts to conduct an extensive analysis of BAs for SMP using open-source datasets.

III. Research Methodology

This section presents a detailed elucidation of the seven empirically collected open-source datasets used in this study and the process of collecting them. The independent and dependent variables chosen for the current study are also described in this section. A careful attempt is made while selecting the independent variables. All the possible and relevant design-related attributes of the OO paradigm, like abstraction, inheritance, complexity, coupling, and cohesion, are covered to sincerely analyze BAs capabilities for SMP. A collection of metrics picked up from different suites proposed by various researchers is selected, including the famous Chidamber & Kemerer (C&K) metrics suite [47]. However, due to some shortcomings encountered in the C&K metrics suite as identified by Malhotra and Chug [48], such as it does not contain any metric to measure the extent of database handling and also its inability to account for the structural complexity that exists in any software; two more metric suites are also considered, Henderson-Sellers [49] and Bansiya & Davis [50]. In totality, a set of seventeen OO metrics covering all the three metric suites has been used while conducting this study, as compiled in Table II.

The dependent variable used here is 'Change,' defined in respect to the number of lines in the source code that were added, deleted, or modified after delivering the final product to the customer. Further, as stated in Section I, there are two types of software attributes, i.e., internal and external. Internal attributes like coupling, cohesion, etc. can directly be measured by the developers during different SDLC stages. In contrast, external attributes like maintainability need to be measured indirectly using the metrics calculated for the internal attributes. In this study, an attempt has been made to measure an external attribute, i.e., maintainability (measured through the dependent variable 'Change') based on internal attributes by finding a correlation between different OO metrics & the dependent variable, developing various SMP models using several different BAs.

The overall implementation for the current study has been performed in Python 3 using Jupyter Notebook 5.7.8 platform. An overview of the research methodology being adopted for this study is depicted in Fig. 2. This section is further subdivided into different sub-sections.

A. Datasets and Data Collection

In this study, seven empirically collected datasets, i.e., Abdera, Ivy, jEdit, jTDS, Log4j, Apache Poi, and Apache Rave from various open-

TABLE II. Independent Variables Used in the Study

Metrics	Definition
WMC (Weighted Methods per Class)	WMC measures the static complexity of all the methods, which is the summation of McCabe's cyclomatic complexity of those methods.
DIT (Depth of Inheritance Tree)	DIT measures a class's position in the inheritance hierarchy, root class having this value equal to zero.
NOC (Number of Children)	The number of direct subclasses of a class is measured using the NOC metric.
CBO (Coupling between Object classes)	The number of classes coupled to a particular class is measured through CBO.
RFC (Response for a Class)	The cardinality of the response set of a class is measured through RFC, which is nothing but the sum of the number of local methods & number of methods called by these methods.
LCOM (Lack of Cohesion in Methods)	The number of disjoint sets of local methods is measured through the LCOM metric.
Ca (Afferent Couplings)	The number of classes that call a particular class is counted by the Ca metric.
Ce (Efferent Couplings)	The number of other classes that are called by a particular class is counted by the Ce metric.
NPM (Number of Public Methods)	The number of public methods of a class is counted by the NPM metric.
LCOM3 (Lack of Cohesion among the Methods of a particular Class)	LCOM3 metric is used to overcome specific disadvantages of the LCOM metric.
LOC (Lines of Code)	The number of code lines, excluding comments & blank lines, is measured using the LOC metric.
DAM (Data Access Metrics)	The ratio of the sum of private & protected methods of a particular class to the total number of attributes defined for that class is calculated by the DAM metric.
MOA (Measure of Aggregation)	The percentage of user-defined data in a particular class is calculated by the MOA metric.
MFA (Measure of Functional Abstraction)	The ratio of inherited methods to total methods in a class is calculated by the MFA metric.
CAM (Cohesion among the Methods of a class)	The similarity between different methods of a particular class is computed by the CAM metric.
IC (Inheritance Coupling)	The number of parent classes to which a particular class is coupled is counted by the IC metric.
AMC (Average Method Complexity)	The average value for McCabe's cyclomatic complexity of all the methods is calculated by the AMC metric.

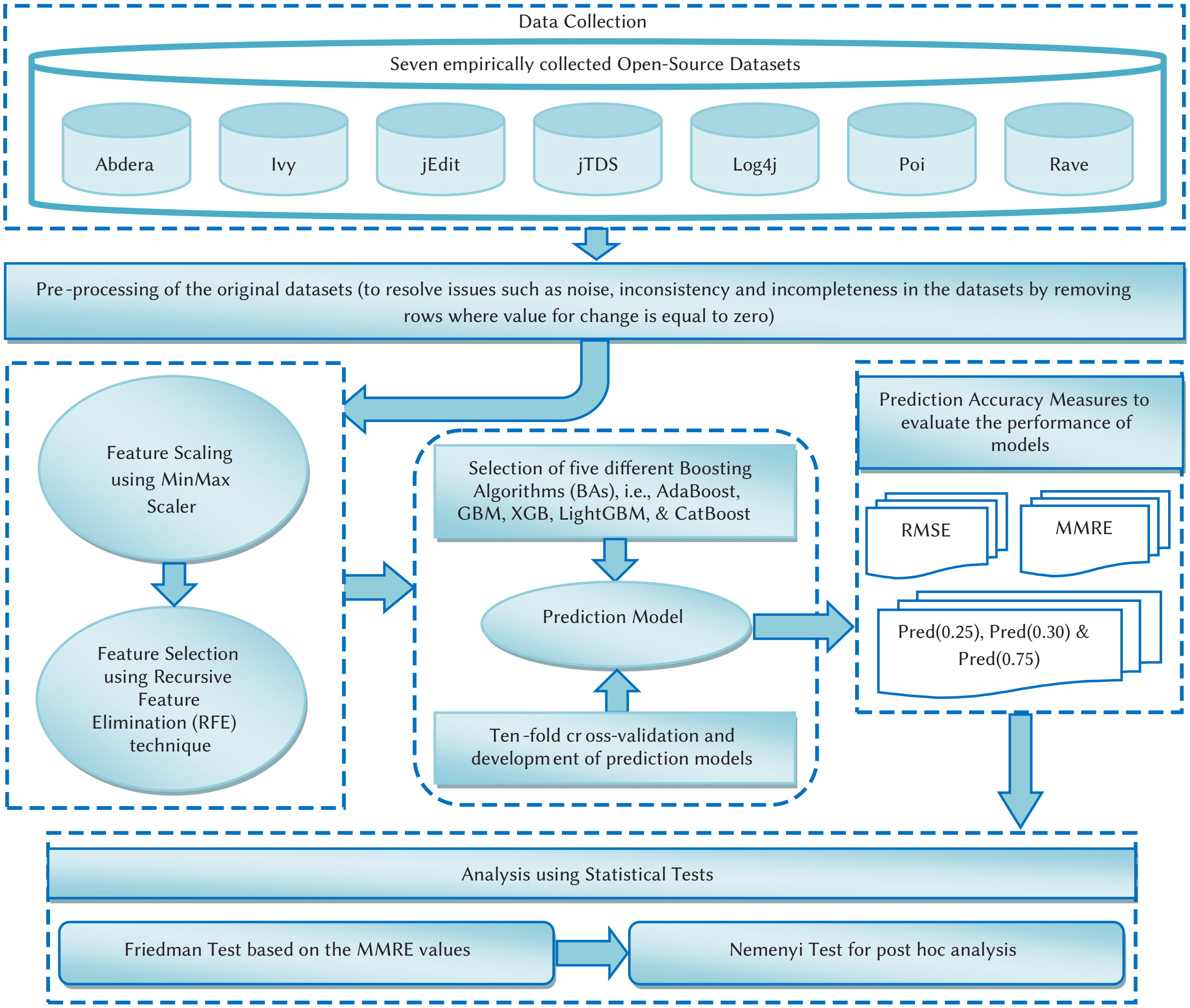

Data Collection
Seven empirically collected Open-Source Datasets
Abdera
Ivy
jEdit
jTDS
Log4j
Poi
Rave
Pre-processing of the original datasets (to resolve issues such as noise, inconsistency and incompleteness in the datasets by removing rows where value for change is equal to zero)
Feature Scaling using MinMax Scaler
Feature Selection using Recursive Feature Elimination (RFE) technique
Selection of five different Boosting Algorithms (BAs), i.e., AdaBoost, GBM, XGB, LightGBM, & CatBoost
Prediction Model
Ten-fold cross-validation and development of prediction models
Prediction Accuracy Measures to evaluate the performance of models
RMSE
MMRE
Pred(0.25), Pred(0.30) & Pred(0.75)
Analysis using Statistical Tests
Friedman Test based on the MMRE values
Nemenyi Test for post hoc analysis

Fig. 2. Research methodology.

source repositories such as SourceForge and GitHub are analyzed using BAs for SMP. The details and description of various datasets follow:

- Abdera (685 classes) - Abdera is an atom parser generator that is used to build functionally a high in performance Internet for both the ends, i.e., the client and the server, by producing such designed documents which are high in quality. (https://github.com/apache/abdera)
- Ivy (613 classes) - Ivy is an assembly of various programs and open source libraries, which allow the broadcast of information using text messages along with a mechanism of subscription, which is usually based on the regular expressions. (https://github.com/apache/ant-ivy)
- jEdit (416 classes) - jEdit is one of the text editors written using Java. It can run on any of the operating systems and is customizable to a great extent. It is also extendable with the help of macros that are written in different scripting languages. (https://sourceforge.net/projects/jedit/)
- jTDS (64 classes) - jTDS is a free and open-source JDBC driver for Sybase ASE & Microsoft SQL Server written purely in Java, which is based on FreeTDS. Also, it is the fastest production-ready JDBC driver that exists currently. (http://jtds.sourceforge.net/)
- Log4j (350 classes) - Log4j is a software that allows control over log statements by the developer to decide which statements can be output having arbitrary granularity. It can entirely be configured at runtime with the help of externally configurable files. (https://github.com/apache/log4j)
- Poi (939 classes) - Poi stands for "Poor Obfuscation Implementation". It is a free open source library written in Java which is used to read and write confusing and hard to interpret document formats of Microsoft Office such as Word, Excel, PowerPoint, etc. (http://poi.apache.org/)
- Rave (671 classes) - Rave is a kind of mash-up supporting different platforms since it is highly customizable. It is a light-weighted and web-based data integration software written in Java that manages various social gadgets by hosting different widgets. It works by combining the data and functionality of two or even more than two sources for creating some new services. (https://rave.apache.org/)

B. Independent Variables

A set of seventeen different OO design metrics taken from different metrics suites proposed by several researchers in their studies [47], [49], [50] has been selected as independent variables of the current study to analyze different BAs for SMP. This set is chosen, keeping in mind that all the essential design-related facets of an OO paradigm like complexity, abstraction, inheritance, coupling, and cohesion are covered. A glimpse of all the selected OO metrics, along with the description, can be viewed in Table II.

C. Dependent Variable

The dependent variable used here is the maintenance effort, which is the 'Change' measured as the number of LOCs for each of the classes that were added, deleted, or modified in the new version compared with the older version of particular software. This comparison is made between two successive versions of the same software where the new version is always the next version, by finding out the common classes of both the versions & subsequently finding the exact count of the lines that have been changed for each class. Each addition or deletion concerning a line is counted as a single change. In contrast, any modification is considered as two changes since, in modification, every deletion is followed by a corresponding addition. Different data points are generated for each class by calculating each of the OO metrics' values and then combining them with the corresponding values of Change made in a particular class.

Further, the details of different open source systems used here, including the version, size (number of classes), and the date of release, are provided in Table III.

TABLE III. Details of Different Open Source Systems Used

Software	Version	Size	Date of Release
Abdera	1.1.2 - 1.1.3	685 classes	15th January 2011 - 21st December 2012
Ivy	2.2.0 - 2.3.0	613 classes	13th June 2012 - 19th August 2015
jEdit	5.1.0 - 5.2.0	416 classes	28th July 2013 - 05th February 2015
jTDS	1.2.8 - 1.3.1	64 classes	08th June 2013
Log4j	1.2.16 - 1.2.17	350 classes	06th April 2010 - 06th May 2012
Poi	3.9 -3.10	939 classes	03rd December 2012 - 08th February 2014
Rave	0.21.1 - 0.22	671 classes	03rd May 2013 - 10th July 2013

D. Pre-processing of the Datasets

Pre-processing is one of the data mining techniques used for transforming raw real-world data into an easy to understand format resolving various issues such as noise (presence of outliers), inconsistency (discrepancy of codes or names), incompleteness (missing attribute values), lack of particular trend and error-proneness in the original datasets. While calculating the dependent variable during this study, a comparison between the old and new versions of all the datasets was made. A Java-based data mining tool for calculating the C&K Java metrics and several other metrics, namely, CKJM extended (http://gromit.iiar.pwr.wroc.pl/p_inf/ckjm/), has been used for empirical data collection. This tool processes the compiled Java files through their byte code & then calculates 19 different size & structure metrics for software. The results in the form of the metrics calculated for each class are displayed on the standard output or are saved in a particular file. The class-wise OO metrics (independent variables) for the older version, for example, for version 2.2.0 of the Ivy dataset, were collected on processing the jar file of that version through CKJM extended tool.

Further, classes common to both the versions, i.e., old and new, were extracted. Those classes added in the new version or deleted from the old version were plainly discarded. Both the library and interface classes were not included in this study. Further, those classes where the value of Change was zero were again excluded while considering the study's datasets. A graphical representation for the percentage reduction achieved for all the seven datasets is provided in Fig. 3.

After pre-processing, comparable classes for both versions were received. Afterward, a line by line comparison of these classes was made with the help of the Beyond Compare tool (https://www.scootersoftware.com/index.php), which provides a quick & easy comparison of files & folders at high speed. It verifies and compares the designated files or folders thoroughly in a byte-by-byte manner and further highlights the specified differences in a different color (generally red). This is required to compute the dependent variable's value, i.e., Change for each class through a line by line comparison. Each addition or deletion of a particular line in a class accounts for a single change, whereas any modification in a particular line of code accounts for two changes, .i.e., a single deletion followed by a single addition.

As shown in Fig. 3, some datasets have even more than 70% of their classes being discarded after pre-processing. However, such systems have been included in the original datasets for this study aiming to include diversified datasets where some datasets have a higher number of classes that get changed between different versions in contrast to the datasets where only a few of the classes get changed in going from one version to another.

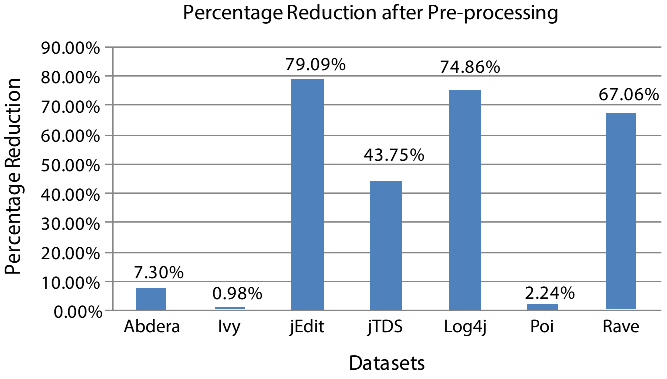


Fig. 3. Percentage reduction after pre-processing.

Further, the descriptive statistics for all the seven datasets have been provided in Table IV. Values highlighted in bold depict the minimum standard deviation for a particular OO metric, whereas the values highlighted in bold with an underline depict the maximum standard deviation for a particular metric.

E. Feature Scaling

Feature Scaling is a technique performed during data pre-processing to standardize a dataset's independent variables in a fixed range. It is also known as data normalization. Feature scaling is done since some algorithms cannot perform appropriately without normalization due to the original datasets' varying range of values. Various feature scaling methods are available for pre-processing data such as MinMaxScaler, RobustScaler, StandardScaler, etc. Of these, MinMaxScaler works by subtracting the minimum value from each of the values in a feature and then dividing it by range where the range is the difference of minimum & maximum values of a feature. In this study also, MinMaxScaler [51] in Python has been applied to normalize all the datasets used here, which transforms all the features by rescaling them to a given range (here, this range is [0, 1]).

F. Feature Selection

Feature selection is a method for choosing a subset of variables or features in ML to develop various models, ensuring the removal of redundant & irrelevant features without incurring the loss of information. It also enhances the prediction models' prediction accuracy since the quality of datasets due to the removal of inconsistent and noisy data and the model's execution time improves. Feature selection algorithms can broadly be classified under two categories: the wrapper methods & the filter methods, as suggested by Kohavi and John [52]. However, a third category is also known as the embedded methods that combine the quality of wrappers and filters and simultaneously perform model fitting and feature selection such as lasso, ridge regression, etc. Wrappers usually evaluate a feature subset's performance based on the learning algorithm's resulting performance, such as forward selection, backward elimination, & RFE, etc. On the other hand, filter methods such as Pearson's correlation, Linear Discriminant Analysis (LDA), etc. generally use some proxy measure for evaluating the importance of the features based on some inherent characteristics without incorporating any learning algorithm. This is contrary to wrapper methods that use error rate for scoring a subset of features. Also, filters are computationally less intensive and faster than wrappers and produce a subset such that it is not tuned to some specific prediction model, making it a more general subset than the one derived from the wrappers.

In this study, one of the wrapper methods called RFE [53] has been used for selecting a subset of independent variables from all the initially selected seventeen independent variables. The improved results obtained in one of the studies conducted for the Intrusion Detection System [54] using the RFE algorithm for ranking the features provided the motivation for using the RFE algorithm in the current study as well for the selection of features. According to the study mentioned above, RFE improves accuracy by counting only the essential features while training, which reduces the learning time. An overall improvement of 0.4% in precision, between 16.2% and 26.8% improvement in false-negative rate, and a one-third reduction in time is achieved. RFE, in general, fits a model by recursively removing the weakest features by taking into account smaller & smaller groups of features, based on the significance of each of the features till a desired count for the features is subsequently reached. This importance is adjudged based on an external estimator, here, therespective boosting algorithms, that assign some weights to the features. Initially, the estimator training is done using a complete set of features, and the importance is obtained for each of the features. After this, the features with the least importance are removed from the initial set (here, one at each iteration since the value for parameter '*step,*' i.e., the count for features to be removed at each iteration is set equal to 1). This procedure is repeated for the reduced set in a recursive manner until a desired set of features that should be selected is eventually obtained. However, in the current study, the default value for the parameter '*n_features_to_select,*' i.e., the count of the features to be selected has been used. This default value is '*None*' and selects half of the total features leaving eight out of seventeen variables in this study, almost equal to half. This way, the RFE algorithm reduces the initial feature set by 52.94%.

Here, only the default values of different parameters for RFE have been used without any changes since the primary focus of the current study is to explore boosting algorithms. However, analyzing the role of tuning of different RFE algorithm parameters for feature selection can form a good base for future studies.

Features selected for all the seven datasets obtained by applying the RFE feature selection algorithm are presented in Table V. It is evident that out of a total of seventeen independent variables, LCOM, NPM, and LOC are found to be the most commonly selected variables. Following them, WMC, RFC, and Ce are the second most commonly selected independent variables. However, DIT, NOC, Ca, and IC have not been selected for any dataset. Also, MOA and MFA came out to be the least significant variables based on the RFE algorithm.

Further, the results in Table V show that each dataset has a different set of features obtained from the RFE algorithm. This difference can be explained through the descriptive statistics presented in Table IV, where values highlighted in bold depict the minimum values, whereas those highlighted in bold with an underline depict the maximum values for a particular metric. The values for standard deviation shows the extent of variation in different OO metrics' values for each of the datasets. The difference in the standard deviations of each of the OO metrics for all the seven datasets accounts for the difference in selecting various features for every dataset using the RFE method since different metrics affect distinct datasets differently while predicting maintainability. Also, due to a comparatively large difference in the variation of values for some OO metrics that have been calculated by finding the range of standard deviation for each metric from Table IV, only a particular set of metrics are selected by the RFE algorithm. For example, the most commonly selected metrics, i.e., LCOM, LOC, RFC, NPM, WMC, and Ce, have considerable variation in their values for almost all the datasets, which have a significant impact in predicting maintainability. Hence, these metrics have been selected by RFE for almost all the datasets.

G. Boosting Algorithms

This sub-section provides an overview of different BAs, i.e., the ensemble of ML algorithms used in the current study to develop various SMP prediction models. A set of five most commonly used BAs, namely AdaBoost, GBM, XGB, LightGBM, and CatBoost, has been applied to identify specific patterns while training each of the seven datasets. These algorithms explore the complex relationship or the correlation among various independent variables & the dependent variable, using the knowledge derived during the training process for making predictions.

1. AdaBoost

AdaBoost is one of the first ensemble boosting techniques proposed by Freund and Schapire [55], [56] to be adapted in practice to solve both regression & classification problems. It works by creating multiple sequential models from poorly performing models, each correcting the previous model's errors to increase the accuracy to build a reliable model ultimately. It is an iterative ensemble technique that generally uses DTs for modeling. However, any ML technique can be used as a base classifier, provided it accepts the weights on the training set. In the current study, the DT regressor has been used as the base estimator while implementing AdaBoost. The basic idea behind AdaBoost is to ensure that the unusual observations are predicted accurately by setting up the weights of classifiers and training data samples in every iteration. AdaBoost is expected to fulfill two main conditions; first, the classifier's interactive training on several weighted training examples should be done, and second, it should try to provide an accurate fit for the above examples in each iteration by minimizing the error in training.

2. GBM

GBM is another ensemble ML algorithm used for classification & regression problems by combining multiple weak learners to develop a strong learner. Friedman described GBM in two of his popular studies in 1999 and 2001 [57], [58]. Generally, Regression Trees (RTs) are used as base learners, and each tree is built subsequently in a series based on the errors measured by the previous tree, and the foremost goal is to overcome these errors. The difference here is that the weights are not incremented for the misclassified values; instead, an attempt is made to optimize and reduce the loss function that adds several weak learners by adding some new model. Broadly, GBM comprises three main components, i.e., the loss function that should be optimized, an additive model for minimizing the loss function & a weak learner for making the predictions.

3. XGB

XGB is a highly effective, novel, and advanced implementation for the GBM ensemble ML algorithm, particularly RTs and K classification.

TABLE IV. Descriptive Statistics for Open Source Datasets (SD = Standard Deviation)

Metric	Abdera			Ivy			jEdit			jTDS		
	Min	Max	SD	Min	Max	SD	Min	Max	SD	Min	Max	SD
WMC	0	255	21.21	1	243	21.64	1	275	34.34	0	211	**42.16**
DIT	0	4	0.63	0	4	0.60	0	7	**1.72**	0	3	0.62
NOC	0	17	1.02	0	17	1.22	0	20	2.22	0	2	0.44
CBO	0	17	1.94	0	17	1.96	1	396	**44.57**	0	34	6.84
RFC	0	256	21.23	2	244	21.64	1	570	**89.92**	0	293	70.72
LCOM	0	32385	1724.93	0	29403	2158.27	0	21943	2605.30	0	21831	**3432.78**
Ca	0	14	1.64	0	17	1.74	0	327	**37.20**	0	30	4.95
Ce	0	5	0.93	0	9	1.17	0	116	**14.86**	0	30	5.07
NPM	0	254	20.65	0	215	18.96	0	228	27.81	0	191	**40.07**
LCOM3	1.0039	2	0.43	1.0041	2	0.42	0	2	0.57	0	2	0.47
LOC	0	1531	123.34	6	1461	132.98	1	10007	**1471.109**	4	8251	1448.33
DAM	0	1	**0.49**	0	1	0.47	0	1	0.44	0	1	**0.39**
MOA	0	327	**14.81**	0	7	0.69	0	13	2.59	0	14	2.40
MFA	0	1	0.17	0	1	0.12	0	0.9987	0.37	0	1	0.23
CAM	0	1	0.30	0.0556	1	0.28	0.0455	1	0.22	0	1	**0.19**
IC	0	3	0.28	0	2	0.21	0	3	**0.58**	0	2	0.41
AMC	0	5	2.24	0	5	2.39	0	139.451	32.32	0	255.11	**46.30**
Change	2	14667	1172.30	2	17586	**1434.72**	1	249	**43.54**	0	355	59.86

Metric	Log4j			Poi			Rave		
	Min	Max	SD	Min	Max	SD	Min	Max	SD
WMC	1	104	13.35	0	165	14.19	0	62	**9.69**
DIT	0	6	1.48	0	5	0.70	0	4	**0.25**
NOC	0	4	0.64	0	151	**5.13**	0	2	**0.21**
CBO	0	76	10.98	0	228	17.25	0	4	**0.69**
RFC	1	130	25.05	0	426	33.42	0	63	**9.71**
LCOM	0	5356	575.44	0	5908	475.71	0	1891	**209.75**
Ca	0	65	9.16	0	228	14.52	0	3	**0.41**
Ce	0	29	4.81	0	167	9.60	0	2	**0.43**
NPM	0	31	**7.50**	0	140	12.51	0	57	9.24
LCOM3	0	2	0.48	0	2	**0.60**	0	2	**0.39**
LOC	3	1864	283.14	0	4455	370.97	0	405	**59.97**
DAM	0	1	0.45	0	1	0.43	0	1	0.43
MOA	0	14	2.13	0	49	3.54	0	3	**0.29**
MFA	0	1	**0.38**	0	1	0.18	0	1	**0.09**
CAM	0.0726	1	0.24	0	1	0.24	0	1	**0.31**
IC	0	3	0.51	0	3	0.27	0	1	**0.20**
AMC	0	205	25.83	0	392.2222	23.40	0	5	**1.69**
Change	1	1612	194.36	2	17956	1331.08	1	470	55.08

It was proposed by Chen and Guestrin in 2016 [30]. It prevents over-fitting and intends towards the optimization of computational resources. It is possible through the simplification of objective functions by allowing the combinations of regularization, provided an optimum computational speed is also maintained alongside. During the training phase, automatic parallel calculations are performed for various functions in XGB. XGB is approximately ten times faster than any other BA and is also known as a "regularized boosting technique." In the current study, 'gbtree' booster, i.e., tree-based models, have been used to run at every iteration.

4. LightGBM

LightGBM was proposed by Ke et al. [59] as a newer implementation of Gradient Boosting DT (GBDT). It is a fast and distributed framework based on DT algorithms and is used for different ML tasks such as ranking & classification. It makes use of a leaf-wise strategy while splitting the trees with the best fit, unlike other algorithms that use . a level-wise or depth-wise approach. Also, LightGBM being leaf-wise is more accurate than other BAs since it can reduce more loss while growing on the same leaf, and it also ensures reduced memory consumption. In the current study, while implementing LightGBM, RF has been used as the basic boosting type.

TABLE V. Feature Selection Through RFE Algorithm

Datasets	Features selected through RFE Algorithm
Abdera	WMC, CBO, RFC, LCOM, NPM, LCOM3, LOC, CAM
Ivy	WMC, RFC, LCOM, Ce, NPM, LCOM3, LOC, CAM
jEdit	RFC, LCOM, Ce, NPM, LCOM3, LOC, MFA, AMC
jTDS	WMC, CBO, RFC, LCOM, Ce, NPM, LOC, AMC
Log4j	WMC, LCOM, Ce, NPM, LOC, DAM, MOA, CAM
Poi	WMC, CBO, RFC, LCOM, Ce, NPM, LOC, CAM
Rave	WMC, RFC, LCOM, Ce, NPM, LOC, DAM, CAM

5. CatBoost

The fundamental algorithmic approaches behind CatBoost were explained by Prokhorenkova et al. [60] in 2018 in one of their studies. CatBoost is also a GBDT for handling categorical features well. It allows one to use the complete dataset for training, and an extensive pre-processing of data is not required here. Rather than the pre-processing time, CatBoost deals with the categorical features while training. As per authors, target statistics can efficiently handle categorical features ensuring minimum loss of information. However, in regression, the initial value is calculated using a standard technique where the average value for a label in the dataset is considered. Overall, CatBoost is a robust, easy-to-use, and high in performance BA in the family of ML algorithms.

H. Cross-Validation Technique

Cross-validation is a model validation technique to account for the accuracy of a prediction model on new data & to check if this model can be generalized for real-world datasets. It is used in a scenario where the ultimate goal is prediction. For prediction, the model is trained using a known dataset (training set), whereas testing of this model is done on a new dataset (test set) after training. Several types of cross-validation approaches exist, such as LOO, *k*-fold, hold-out, etc. However, the *k*-fold technique is one of the most basic cross-validation forms with *k* equal to 10 [61]. Its significance lies in the fact that it can use the dataset for dual purpose, i.e., training and testing. As per literature, 5-fold and 10-fold approaches are the most commonly used cross-validation approaches to design a model. According to [61], *k*-fold validation using moderate values of k, i.e., between 10-20, helps minimize the variance with an increase in the bias; *k* equal to 10 being the most preferred, most frequently used, and the most recommended one.

Further, if the value of *k* decreases say between 2-5, along with smaller sample size, then variance seeks in because of the instability in the training set, which further increases the variance. Hence, the 10-fold cross-validation technique has been selected for the current study, keeping the above points in mind. In this technique, the complete dataset is sub-divided into 10 equal partitions, of which one partition is considered validation data for testing the model, whereas the rest of the partitions are utilized to train the prediction model. The same process is iteratively repeated 10 times for each of the 10 partitions, each partition being used as the validation set exactly once. Lastly, a single final estimation is reached by calculating an average of the 10 results obtained above.

I. Prediction Accuracy Measures

This section presents various prediction accuracy measures for the current study to assess various BAs performance for SMP for all the seven datasets. Estimating and assessing the accuracy of a prediction model is an essential part of any study. This is done by comparing the dependent variable's predicted value with its actual value and finding the corresponding value of the error. In literature, different residual-based prediction accuracy measures have been suggested by various researchers. However, in this study, the following three measures of accuracy have been selected for estimating the accuracy of the proposed prediction models, as suggested by Conte et al. [62] and Kitchenham et al. [63], [64] in their studies.

1. Root Mean Square Error (RMSE)

RMSE is the measure of standard deviation in the prediction errors, i.e., the residuals. It is calculated by taking the square root of Mean Square Error (MSE) using the formula defined by Conte et al. (1986) [62].

$$MSE = \frac{1}{n}\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2 \quad (1)$$

where y_i is the i^{th} value being observed & $\hat{y}_i$ is the i^{th} value, which is predicted by the prediction model. Further, from this formula, the formula for RMSE can also be derived.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2} \quad (2)$$

Residuals measure the data points' remoteness from the standard regression line, and RMSE measures these residuals' spread. Alternately, RMSE describes the concentration of the data around the best line of fit. RMSE is one of the most widely and frequently used indicators to account for the goodness of fit in regression models. The significance of using RMSE may be attributed to the risk-averse predictors, where large deviations are penalized more in comparison to small deviations by RMSE. This happens as RMSE is based on the mean or average value obtained by summing the residuals' squared values. Another implication of squaring the errors in determining the importance of RMSE lies in the fact that RMSE would be even more useful in cases where large errors may particularly be highly undesirable. RMSE is highly useful to compare several prediction models developed using different techniques. The magnitudes of the prediction errors for several times are accumulated into a compound indicator of the predictive power with the help of RMSE. Further, RMSE values may range from 0 to ∞ and are inconsiderate towards the errors' direction. RMSE is a negatively-oriented indicator in which lower values are considered to be the better values. Also, RMSE does not essentially increase with a rise in the error variance but increases with a rise in the variance related to the frequency distribution of the magnitude of the errors.

2. Mean Magnitude of Relative Error (MMRE)

MMRE is the most frequently used quality indicator in software engineering while accounting for the performance of various software estimation models defined by Conte et al. (1986) [62].

$$MMRE = \frac{1}{n}\sum_{i=1}^{i=n}\left(\frac{|y_i - \hat{y}_i|}{y_i}\right) \quad (3)$$

MMRE is different from relative error in the sense that unlike relative error, the absolute value of the differences between the actual & predicted values is used while calculating MMRE. The use of absolute value prevents both underestimation and overestimation by canceling each other out. MMRE measures the variance or the spread of accuracy (predicted / actual). To better understand what MMRE measures, 'y' is considered a normally distributed random variable having μ and σ^2 as the mean and variance, respectively. Iglewicz [65] has already illustrated the following for a sample of 'n' observations, where $\bar{y}$ is the mean of those observations:

$$d_n = \frac{1}{n}\sum|y_i - \bar{y}| \rightarrow \sigma\sqrt{\frac{\pi}{2}} \text{ as } n \rightarrow \infty \quad (4)$$

On re-writing MMRE as below:

$$\frac{1}{n}\sum_{i=1}^{i=n}\left|\frac{\hat{y}_i}{y_i} - 1\right| = \frac{1}{n}\sum_{i=n}^{i=1}|z_i - 1| \quad (5)$$

it is evident that if $\hat{y}_i$ is an unbiased estimation of y_i, then the value of $z_i = \frac{\hat{y}_i}{y_i}$ as expected is equal to 1. If z_i is normally distributed having mean and variance equal to 1 and σ_z, respectively, then MMRE tends towards the value of $\sigma_z\frac{\pi}{2}$. This illustrates that MMRE estimates the spread or the variance of the 'z' variable, which is not that susceptible to the large outliers as the RMS estimate is. As MMRE measures the spread, it would be wrong to call it a prediction accuracy metric. However, 'z' has an optimal defined value equal to 1, indicating whether or not the prediction system's estimation is under or overestimated and hence a better criterion to indicate the prediction accuracy. This further shows that any prediction model's quality can be described in respect of the average value of the 'z' variable, and MMRE is used for assessing the variability of this variable 'z.'

3. Pred(m)

It measures the proportion of the values predicted by the prediction models with a magnitude of MRE lower than or equal to a specific value.

$$Pr\,e\,d(m) = \frac{k}{n} \tag{6}$$

where ‘*m*’ represents the particular specified value, ‘*k*’ refers to the number of predictions in the dataset whose MRE is lower than or equal to ‘*m*’ & ‘*n*’ is the total number of observations in the dataset. Pred(*m*) measures the kurtosis or the shape of the accuracy (predicted / actual). Pred(*m*) is the percentage of predictions within *m*% of the initial or the actual values. ‘*m*’ is usually set to 25, 30, & 75 such that Pred(*m*) shows how much proportion of the predictions lie within the tolerance of 25%, 30%, & 75% respectively. Further, Pred(*m*) is inconsiderate to the extent of the inaccuracy of the predictions that lie beyond a particular level of tolerance. For example, for two different prediction models whose predictions deviate by 27% and 270%, respectively; a Pred(25) indicator will not differentiate between the two models. Like MMRE, Pred(*m*) should preferably be formulated for prediction by considering the actuals' percentage lying within *m*% of the prediction. As mentioned earlier, Pred(*m*) is a measure for kurtosis that provides the degree to which a particular distribution has been peaked surrounding the central value. To better understand Pred(*m*), consider a case where certain distribution is more peaked than a normal distribution. As a result, if a sample is selected from a distribution having more peak, then it would have comparatively more values within 25% (in case of Pred(25)) of the mean value than normal. On the other hand, if a sample is selected from a distribution having a flatter peak, it would have comparatively fewer values within 25% than the normal scenario.

J. Friedman Test for Ranking the Performance

Friedman test [66] is a statistical test, which is non-parametric in nature, to rank the performance of various algorithms used in a study by finding any significant difference between those algorithms' performance. Here, this test has been used to rank the performance of different BAs used in this study. Before the test, a hypothesis is formulated.

Null Hypothesis (H_0) - No significant difference exists between the performances of various BAs used in this study.

Alternate Hypothesis (H_1) - A significant difference exists between the performances of various BAs used in this study.

Further, the Friedman measure is calculated using the given formula.

$$\chi_r^2 = \left(\frac{12}{Nk(k+1)}\sum_{i=1}^{k} R^2\right) - 3N(k+1) \tag{7}$$

where *R* represents the average rank for each BA, *N* stands for the number of datasets used in the current study, and *k* represents the number of BAs considered for the ranking. The value for $\chi_{calculated}$ is calculated using (7) and further compared with $\chi_{tabulated}$ given in the distribution table for chi-square. If $\chi_{calculated}$, which is the Friedman measure, falls in the critical region, it is concluded that a significant difference exists between the performance of various BAs, thereby rejecting the null hypothesis & accepting the alternate hypothesis. However, if $\chi_{calculated}$ does not fall in the critical region, it is then concluded that no significant difference exists between the performance of various BAs, thereby rejecting the alternate hypothesis and accepting the null hypothesis.

Each BA is ranked individually with the help of Friedman's Individual Rank (FIR) using (8).

$$FIR = \frac{C}{N} \tag{8}$$

where *C* represents the cumulative rank & *N* represents the total number of datasets. Based on the FIR values calculated for each BA, one having the lowest FIR value is declared to be the best performer whereas, on the other hand, BA having the highest FIR value is declared to be the worst performer. Further, suppose the values of FIR obtained for various prediction accuracy measures are found significant. In that case, post hoc analysis should be done using the Nemenyi test to check if the difference between various mean ranks obtained by the Friedman test is statistically significant or not. However, in this study, both Friedman and Nemenyi tests are performed only for MMRE.

K. Nemenyi Test for Post Hoc Analysis

Nemenyi test [67] is a test in statistics for post hoc analysis that intends to find groups of data that differ when statistical tests such as the Friedman test for multiple comparisons rejects the null hypothesis stating that no significant difference exists between the performance of various groups of data. This test is used to make pair-wise tests of performance for comparing the performance of various BAs used in this study to find if any statistical difference exists among them. The first step for conducting the Nemenyi test is to calculate the Critical Difference (CD), which depends on the total number of BAs & the number of datasets used, along with the level of significance, using (9).

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{9}$$

where *k* represents the total number of BAs, *N* represents the number of data samples & q_α is the critical value as suggested by Demsar [68] in his study; based on the Studentized range statistics for a particular significance level. After calculating CD, the individual differences between the FIR values of different pairs of BAs are calculated to compare each possible pair of BAs' performance during the post hoc analysis. If the difference calculated for each possible pair of BAs comes out to be either more than or equal to CD, then the performance of that particular pair is considered statistically significant for the selected level of significance. On the other hand, if this difference is less than CD, then that particular pair's performance is statistically not significant.

IV. Results & Discussions

This section presents the results of the current study & a detailed discussion and analysis of these results to analyze different BAs for SMP using open-source datasets. A few of the selected plots for all the seven datasets showing true versus predicted values for the best performing BA based on MMRE values are presented in Fig. 4.

It is noted that based on the MMRE values, XGB performed the best for six out of the seven datasets except for jEdit, for which CatBoost performed the best. Subsequently, various RQs framed for the current study in the introduction section are answered in this section.

RQ1: Whether BAs can be applied for SMP?

Various prediction accuracy measures, i.e., RMSE, MMRE, and Pred(0.25), Pred(0.30) & Pred(0.75), have been used to analyze the performance of different BAs used in this study for all the seven datasets using (2), (3) and (6), respectively. The results obtained for all the five BAs validated using ten-fold cross-validation are presented, compared, and analyzed in this section.

Table VI provides the RMSE values for each of the BAs for all the seven datasets. The best value of RMSE for each dataset is marked in bold. It is clear from Table VI that based on the RMSE, GBM performed the best, resulting in the lowest RMSE values for three of the seven datasets, namely jEdit, jTDS, and Log4j, i.e., for 42.86% of the total datasets. Similarly, LightGBM performed the second-best in terms of RMSE for Ivy and Poi, i.e., for 28.57% of the datasets, whereas AdaBoost and CatBoost performed well for Abdera and Rave, respectively.

However, XGB came out to be the worst performer since it did not perform well for any of the datasets when RMSE is considered the accuracy measure. Overall, if we look at Table VI, it is concluded that the best RMSE value equal to 43.42 is obtained for the jEdit dataset using GBM.

TABLE VI. RMSE Values For All the Seven Datasets using BAs

Accuracy Measure	**RMSE**						
Datasets	Abdera	Ivy	jEdit	jTDS	Log4j	Poi	Rave
Boosting Algorithm							
AdaBoost	**986.25**	1251.82	46.63	70.15	184.08	1241.31	55.81
GBM	1150.49	1389.05	**43.42**	**67.54**	**179.01**	1310.24	55.30
XGB	1164.34	1368.90	45.81	76.14	211.25	1324.29	58.57
LightGBM	1070.95	**1199.34**	54.92	70.21	195.85	**1197.91**	75.98
CatBoost	1077.89	1278.42	44.24	76.03	201.73	1230.27	**54.78**

The MMRE values for each of the BAs validated using ten-fold cross-validation for all the datasets are provided in Table VII. Also, the least obtained values of MMRE for each dataset, which are also the best, are marked in bold since a low value of MMRE indicates less error in prediction and hence better accuracy. Every row shows MMRE values for a particular BA on a specific dataset.

TABLE VII. MMRE Values For All the Seven Datasets using BAs

Accuracy Measure	**MMRE**						
Datasets	Abdera	Ivy	jEdit	jTDS	Log4j	Poi	Rave
Boosting Algorithm							
AdaBoost	4.36	7.31	2.22	2.04	7.98	6.49	2.81
GBM	4.56	7.87	3.86	2.97	8.75	6.44	3.91
XGB	**1.84**	**2.98**	1.77	**0.90**	**3.82**	**2.85**	**1.43**
LightGBM	4.91	7.45	4.94	3.33	8.90	6.88	5.01
CatBoost	2.93	4.62	**1.71**	2.54	6.26	4.09	1.92

It is concluded from Table VII that based on the MMRE values so obtained, XGB performed the best for six out of seven datasets, i.e., for 85.71% of the total datasets by providing the least values for MMRE. However, in the jEdit dataset, CatBoost performed the best in terms of MMRE with a value equal to 1.71. Overall, XGB performed the best with the jTDS dataset having MMRE value equal to 0.90 when MMRE is considered an accuracy measure to analyze different BAs performance over seven open-source datasets.

Further, the prediction accuracy of all the BAs for each of the datasets has been calculated at 25%, 30% & 75%, and results are summed up in Table VIII where each column for a particular dataset is further subdivided into three columns; one each for Pred(0.25), Pred(0.30) & Pred(0.75). Best obtained values are highlighted in the table for each dataset & each prediction accuracy level, i.e., 25%, 30%, and 75%.

On analyzing the values in Table VIII, it is observed that for Pred(0.25), which ranges up to 31%, CatBoost BA is found to be the most accurate in the case of Abdera. If we consider Pred(0.30) for prediction accuracy, which ranges up to 36%, it is found that CatBoost for Abdera performed the best. Again, for Pred(0.75), which ranges up to 79%, it is observed that XGB performed the best for Abdera by providing the highest prediction accuracy equal to 79%, which further assures the effectiveness of BAs for SMP. Overall, it is concluded that the best prediction accuracies are obtained for Abdera. Also, in the case of Pred(0.30), LightGBM performed the best for three out of the seven datasets, i.e., Ivy, jEdit, and Poi, whereas, in the case of Pred(0.75), XGB gave the best performance for six out of the seven datasets (excluding Log4j) with prediction accuracies ranging from 51% to 79% which are satisfactory and reasonable.

Further, the difference in the performance of different BAs for different datasets can be accounted to the wide range and variations in the values of various OO metrics and the dependent variable 'Change' measured through standard deviation as presented in Table IV, representing the descriptive statistics for each of the seven datasets. Hence, a different range of values is obtained for various prediction accuracy measures used in the current study. For example, if we consider the predictor variable 'Change,' then the difference between the maximum and minimum values for standard deviation so obtained or the range for standard deviation is 1391.18, which is really wide and hence the difference in performance.

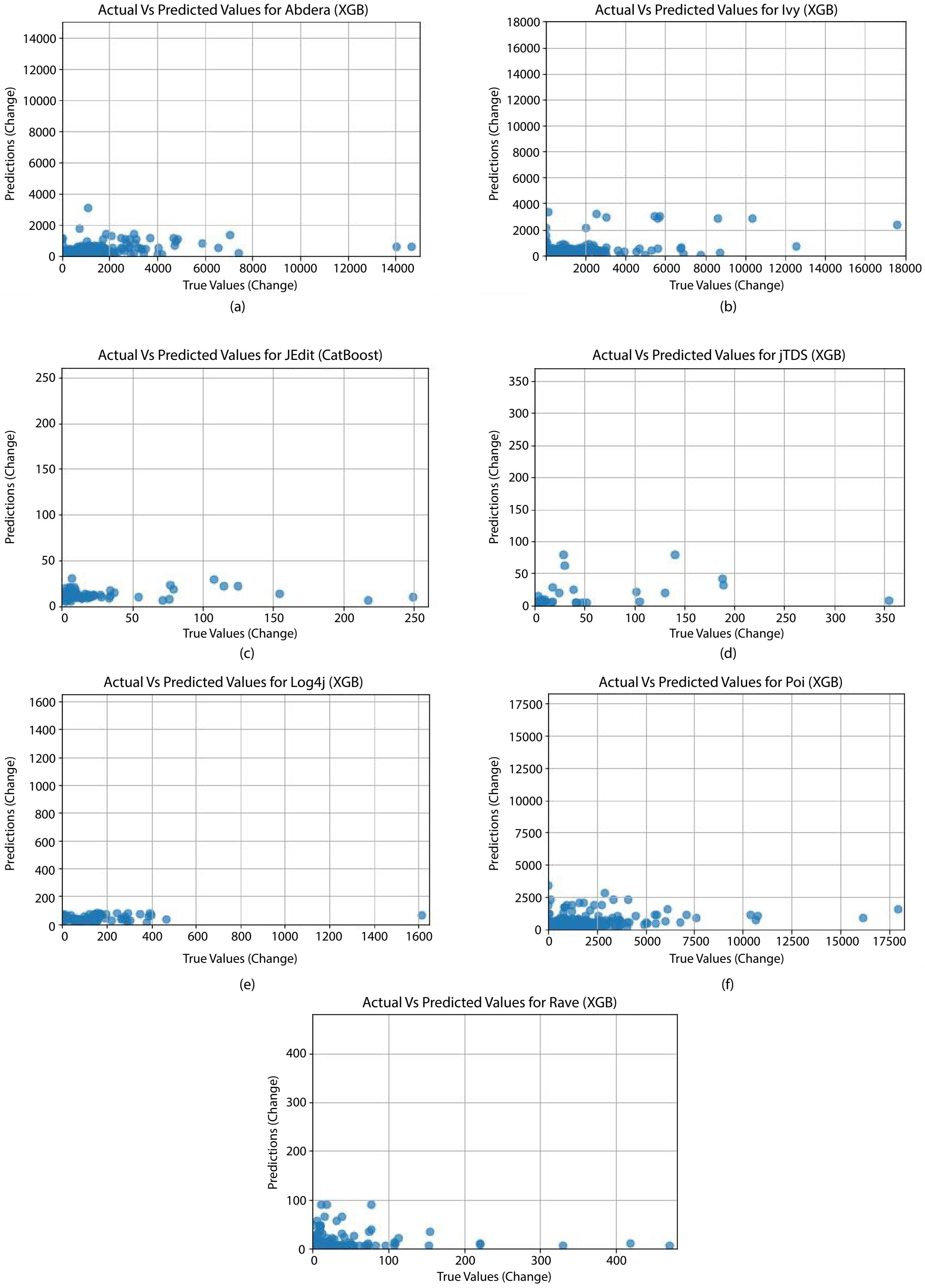
Hence, based on each of the BAs' overall performance for all the seven datasets based on the values obtained for the five prediction accuracy measures and from the comparative analysis of these measures, it can be concluded that BAs can effectively be applied for SMP.

RQ2: Which BA performs the best amongst different BAs based on various prediction accuracy measures for different open-source datasets?

A non-parametric statistical test named the Friedman test is applied for an extensive analysis of different BAs used in the current study to determine if a significant difference exists between various BAs. Friedman's test is selected because it is a non-parametric test; it is safe and robust as it does not assume homogeneity of variance or the normal distributions as recommended by Demsar in his work [68]. The Friedman test has been conducted for comparing the performance of five different BAs applied on seven different datasets based on the MMRE values by calculating the value of critical region for the level of significance equal to 5% & degree of freedom equal to 4, i.e., 5 BAs minus 1 (or *k*-1 where '*k*' is the total number of BAs used in this study). Value for $\chi_{tabulated}$ is read from the Chi-square table corresponding to the 95% significance level and degree of freedom equal to 4.

TABLE VIII. Pred(*m*) Values for All the Seven Datasets Using BAs

Accuracy Measure	**Pred (m)**																				
Datasets	Abdera			Ivy			jEdit			jTDS			Log4j			Poi			Rave		
Boosting Algorithm	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)
AdaBoost	0.24	0.28	0.62	**0.21**	0.24	0.54	0.20	0.22	0.51	0.14	**0.19**	0.33	0.25	0.28	**0.64**	**0.24**	**0.27**	0.57	0.12	0.14	0.47
GBM	0.15	0.19	0.41	0.12	0.16	0.39	0.09	0.13	0.30	0.14	**0.19**	0.44	**0.27**	**0.30**	0.60	0.13	0.16	0.45	0.13	0.16	0.36
XGB	0.21	0.25	**0.79**	0.17	0.21	**0.66**	0.07	0.14	**0.52**	0.14	0.14	**0.53**	0.02	0.02	0.49	0.13	0.17	**0.66**	**0.14**	**0.19**	**0.51**
LightGBM	0.23	0.27	0.57	**0.21**	**0.26**	0.53	**0.21**	**0.23**	0.47	**0.17**	0.17	0.36	0.24	0.25	0.59	0.23	**0.27**	0.57	0.12	0.14	0.39
CatBoost	**0.31**	**0.36**	0.76	**0.21**	0.24	0.61	0.11	0.14	0.44	0.08	0.11	0.44	0.07	0.11	**0.64**	0.21	0.24	**0.66**	**0.14**	0.17	0.49



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 4. Plots for (a) Abdera; (b) Ivy; (c) jEdit; (d) jTDS; (e) Log4j; (f) Poi; (g) Rave showing true versus predicted values for the best performing BA based on the MMRE values.

According to the null hypothesis of Friedman's test, which states that no significant difference exists between the performance of various BAs used in this study, it was found that at 0.05 level of significance, $\chi_{calculated}$ which is the Friedman measure lies in the critical region for MMRE. Therefore, it is concluded by accepting the alternative hypothesis and rejecting the null hypothesis that a significant difference exists between the performances of various BAs used in this study. Test statistics for the Friedman test are stated in Table IX.

TABLE IX. Friedman Test - Test Statistics for MMRE

N	7
Chi-Square	24.914
Df	4
Asymp. Sig.	.000

Further, each BA is ranked for its performance by calculating FIR from (8) based on MMRE, and the values obtained for the mean ranks of different BAs for MMRE are compiled in Table X. It is also known that the lowest mean rank indicates the best performance. Hence, it is evident from Table X that based on the MMRE values, XGB performed the best, whereas CatBoost performed the second best. Also, LightGBM is found to be the worst performer.

TABLE X. Mean Ranking of BAs on Applying Friedman Test for MMRE

Boosting Algorithm	XGB	CatBoost	AdaBoost	GBM	LightGBM
Mean Rank	1.14	2.00	3.00	4.00	4.86

On exploring the reason for the difference in the performances of XGB and LightGBM, it was found that due to the use of Newton boosting in XGB, it is likely to learn better structures. Apart from this, XGB consists of an extra parameter for regularization, namely column sub-sampling (including built-in L1 & L2 regularization that prevents the model from being over-fitted) for reducing the correlation between each of the trees further. Also, XGB uses a histogram-based pre-sorted algorithm for computing the best split and achieve faster training. In contrast, LightGBM uses the GOSS technique, i.e., Gradient-based One Side Sampling, for filtering the data samples to find a value for the split. Unlike other algorithms, where trees grow horizontally (level wise), in LightGBM, trees grow vertically (leaf wise) by choosing the leaf having maximum delta loss.

A further implication of the results can be the utilization of boosting algorithms, especially the XGB, for developing different prediction models in a scenario where training data is limited, time for training is less, and the expertise for tuning of parameters also lacks.

RQ3: What is the comparative performance of various BAs during post hoc analysis when MMRE is taken as an accuracy measure?

After the Friedman test, post hoc analysis using the Nemenyi test is performed to check if the differences between the performances of various BAs based on the FIR values, as concluded in RQ2 above, are statistically significant or not.

The value for CD is calculated to be equal to 2.31 using (9) where k is taken to be 5 (number of BAs), and N is taken to be 7 (number of datasets). After this, all the possible pairs of BAs are formed with every other BA for calculating the rank differences between them, i.e., between the FIR values so obtained. Here, ten such combinations are formed for five different BAs for MMRE, and the same results are compiled in Table XI.

Values for differences in ranks greater than or equal to CD, i.e., 2.31, are shown in bold in Table XI. It is observed that 3 out of 10, i.e., 30% of the total pairs of BAs have been highlighted, which means 30% of the pairs have the difference above or equal to CD, showing that the performance of these pairs is found to be significantly different using Nemenyi test.

Differences calculated in Table XI also show that XGB performed better than GBM, and LightGBM, whereas CatBoost performed better than LightGBM only. Therefore, from this post hoc analysis of MMRE values, it is concluded that XGB and CatBoost significantly outperformed the rest of the BAs. However, the differences between the performances of all other pairs of BAs have not been found significant.

TABLE XI. Pair-Wise Rank Differences Between Different BAs in Terms of MMRE

Boosting Algorithm	XGB	CatBoost	AdaBoost	GBM	LightGBM
XGB	-	0.86	1.86	**2.86**	**3.72**
CatBoost		-	1.00	2.00	**2.86**
AdaBoost			-	1.00	1.86
GBM				-	0.86
LightGBM					-

RQ4: What is the comparison between the results obtained on applying various ML algorithms (other than the BAs) and the results obtained on implementing BAs?

To show why BAs are so good compared to other ML algorithms, a comparison of results between different ML algorithms and the BAs has been made through the RQ mentioned above based on the RMSE, MMRE, and different Pred(m) values. Four different ML algorithms belonging to four different categories, i.e., tree-based models (DTs), neural network-based models (MLP), ensemble models (Bagging), and linear models (Elastic - Net) have been selected for carrying out this comparison. All the four models, along with a brief description, have been presented as follows:

- Decision Trees (DTs): DTs are a supervised and non-parametric ML algorithm for solving classification & regression problems. DTs' primary goal is to develop such predictive models where the response variable is predicted using the knowledge learned from various decision rules that have been inferred through the data attributes. Here, the rules are generated by breaking down the complex process of decision making into several simple decision rules which often provide us with easily interpretable solutions resembling the desired set of solutions [69]. DTs have several advantages, including DTs are easy to understand & interpret, require little or no data preparation, the computational cost is logarithmic to the number of training data points used in the tree, etc.
- Multi-layer Perceptron (MLP): MLP is again a supervised and neural network-based ML algorithm which learns the following function through training on the dataset,

$$f(\bullet): R^{in} \rightarrow R^{out} \tag{10}$$

where 'in' corresponds to the number of input dimensions and 'out' represents the number of output dimensions. For a given set of attributes, say, $X = x_1, x_2, \cdots, x_l$ and a response variable y, MLP can provide a non-linear approximation of the function for regression or a classification problem. MLP consists of one or more hidden non-linear layers between the two layers, i.e., input & output layer. MLPs are capable of learning in real-time, and they can learn non-linear models also. Particularly, in the case of regression, backpropagation has been used for implementing MLP with identity function being the activation function or having no activation function at all for the output layer [70]. Also, the square error is used as the loss function having the response variable as a collection of several continuous values.

- Bagging: Bagging is one of the ensemble ML methods that works by combining the predictions obtained from various base estimators built using a particular ML algorithm to improve the generalizability or robustness of the single estimator. Bagging belongs to the family of averaging methods out of the two prominent families of ensemble methods, i.e., the averaging methods and the boosting methods. The basic idea behind bagging [71] is to implement several independent base estimators (e.g., DTs, MLPs, etc.) over random subdivisions of the initial training set in the first instance and then taking out the average of each of the predictions to obtain a final prediction. Overall, the combined or aggregated bagging estimator is supposed to be better than the single estimators since the variance has been reduced.
- Elastic - Net (EN): EN [72] is a regularized linear ML algorithm for regression, which combines the penalties of two other linear models, i.e., the lasso & the ridge models, in a linear manner having *L1* & *L2* regularization, respectively. This aggregation encourages an efficient learning procedure, especially for the models having few non-zero weights like the lasso, with simultaneous maintenance of the properties of regularization for the ridge method. EN is advantageous in the case of multiple attributes being correlated to each other. However, the lasso is expected to select only one of them, that too randomly, whereas EN is expected to select both of them.

Further, all the ML models mentioned above have been implemented using similar procedures while implementing different BAs. All the algorithms have been implemented for seven open-source datasets (Abdera, Ivy, jEdit, jTDS, Log4j, Poi, & Rave) after pre-processing. Further, feature selection using the RFE algorithm and ten-fold cross-validation has also been performed. The performance of these models has been evaluated using the same performance measures, viz, RMSE, MMRE, Pred(0.25), Pred(0.30), & Pred(0.75) for comparison of the results so obtained with various prediction models that have been built using BAs. The results obtained on applying these four algorithms, i.e., DT, MLP, Bagging, and EN for all the datasets based on RMSE, MMRE, and Pred(m), have been provided in Tables XII, XIII, & XIV, respectively.

TABLE XII. RMSE Values for All the Datasets Using ML Algorithms

Accuracy Measure	**RMSE**						
Datasets	Abdera	Ivy	jEdit	jTDS	Log4j	Poi	Rave
ML Algorithm							
DT	1312.59	1719.58	71.55	**71.60**	198.66	1525.33	79.56
MLP	1203.93	1467.22	48.91	80.64	198.09	**1334.38**	61.76
Bagging	1215.03	1477.21	49.75	79.32	199.28	1452.23	61.72
Elastic - Net	**1174.68**	**1431.94**	**43.27**	73.84	**197.56**	1339.19	**55.29**

TABLE XIII. MMRE Values for All the Datasets Using ML Algorithms

Accuracy Measure	**MMRE**						
Datasets	Abdera	Ivy	jEdit	jTDS	Log4j	Poi	Rave
ML Algorithm							
DT	6.80	15.27	6.19	**0.98**	**9.00**	11.35	**3.15**
MLP	6.14	11.25	7.82	11.19	9.77	6.59	8.33
Bagging	6.58	11.75	8.12	10.43	10.25	11.03	8.20
Elastic - Net	**4.53**	**7.94**	**3.71**	6.33	11.36	**6.37**	3.92

On comparing Table VI and Table XII showing the RMSE values for seven different datasets using BAs and other ML algorithms, respectively, it is observed that BAs have performed better than the other ML algorithms. Precisely, the RMSE values obtained for Abdera, Ivy, and Poi datasets are lower and better using any of the five BAs compared to all the four other ML algorithms, i.e., DT, MLP, bagging, and EN. In the jEdit dataset, four out of the five BAs, i.e., AdaBoost, GBM, XGB, & CatBoost, performed better than three out of the four other ML algorithms, i.e., DT, MLP, & bagging in terms of RMSE. Further, for jTDS and Log4j datasets, three out of the five BAs, viz., AdaBoost, GBM, & LightGBM (i.e., 60% of the total BAs) show comparatively lower values of RMSE than all other ML algorithms. Lastly, in the Rave dataset, CatBoost BA outperformed all the other ML models with a lower value of RMSE, whereas AdaBoost, GBM, & XGB BAs outperformed DT, MLP, & bagging models. Overall, based on the RMSE values provided in Table VI and Table XII and on comparing the lowest RMSE values (values marked in bold) computed for each dataset in both the tables, BAs show a better performance since the lowest, and hence the best RMSE values have been obtained using BAs as compared to other ML algorithms for six (Abdera, Ivy, jTDS, Log4j, Poi, & Rave) out of the seven datasets, i.e., for 85.71% of the datasets. As an example, the least RMSE value equal to 71.60 obtained for the jTDS dataset on applying the DT algorithm reduces to 67.54 on applying GBM BA, leading to an improvement of 5.67%. Subsequently, on analyzing the mean RMSE values obtained for all the BAs taken together and also for all the other ML algorithms as shown in Fig. 5, it is concluded that the performance of BAs (having comparatively lower RMSE values) is better than other ML algorithms for all the seven datasets. An overall improvement in the mean RMSE values equal to 11.14%, 14.86%, 11.94%, 5.68%, 2.03%, 10.76%, and 6.95% for Abdera, Ivy, jEdit, jTDS, Log4j, Poi, and Rave datasets, respectively, has been achieved on applying BAs when compared to other ML algorithms.

Further, based on the MMRE values obtained for BAs and other ML algorithms presented in Table VII and Table XIII, it is evident that BAs performance is undoubtedly better than the other ML algorithms. Specifically, the lower MMRE values for three (AdaBoost, XGB, & CatBoost) out of the five BAs are better than all the four other ML algorithms, i.e., DT, MLP, bagging, & EN for both Abdera and jEdit

TABLE XIV. Pred(*m*) Values for All the Datasets Using ML Algorithms

Accuracy Measure	**Pred (m)**																				
Datasets	Abdera			Ivy			jEdit			jTDS			Log4j			Poi			Rave		
ML Algorithm	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)	(0.25)	(0.30)	(0.75)
DT	**0.20**	**0.25**	**0.62**	**0.17**	**0.20**	**0.54**	**0.20**	**0.21**	**0.52**	**0.31**	**0.31**	**0.61**	**0.24**	**0.26**	0.48	**0.25**	**0.28**	**0.59**	**0.20**	**0.23**	**0.46**
MLP	0.11	0.14	0.33	0.09	0.11	0.28	0.02	0.02	0.17	0.00	0.06	0.22	0.23	**0.26**	**0.61**	0.15	0.19	0.49	0.07	0.09	0.20
Bagging	0.10	0.12	0.30	0.09	0.11	0.26	0.01	0.01	0.15	0.00	0.06	0.25	0.22	0.24	0.59	0.10	0.12	0.28	0.07	0.09	0.20
Elastic - Net	0.14	0.17	0.39	0.11	0.14	0.38	0.10	0.14	0.30	0.14	0.14	0.28	0.19	0.25	0.57	0.13	0.16	0.43	0.11	0.15	0.36

datasets. Also, GBM and LightGBM BAs performed better than three (DT, MLP, & bagging) out of the four other ML algorithms for Abdera and jEdit datasets. Next, considering the Ivy and Log4j datasets, it is observed that lower and better MMRE values have been achieved for all the five BAs as compared to any of the other ML algorithms. In the jTDS dataset, XGB outperformed all other ML algorithms, whereas, rest of the four BAs outperformed three (MLP, bagging, & EN) out of the four other ML algorithms. Proceeding to the Poi dataset, XGB and CatBoost BAs provide lower MMRE values than any other ML algorithms. At the same time, AdaBoost and GBM BAs performed better than three (DT, MLP, & bagging) of the other ML algorithms. Lastly, on considering the Rave dataset, it is found that three out of the five BAs, viz. AdaBoost, XGB, & CatBoost BAs show better MMRE values than all other ML algorithms. However, GBM BA shows better performance than MLP and bagging algorithms, and it performs almost as good as the EN algorithm. Overall, on comparing the lowest MMRE values (values marked in bold) provided for each dataset in Table VII and Table XIII using BAs and other ML algorithms, respectively, it is observed that BAs showcase a better performance due to the lowest and the best-obtained MMRE values for all the seven open-source datasets, .i.e. for 100% of the datasets. As an example, the least MMRE value equal to 9.00 obtained for the Log4j dataset on applying the DT algorithm reduces to 3.82 on applying XGB BA, leading to an improvement of 57.56%. Not only this, the mean MMRE values calculated for all the BAs taken together and for all the other ML algorithms have been depicted in Fig. 6. It is evident from Fig. 6 that lower MMRE values have been obtained using BAs for all the seven datasets considered in this study which, further strengthens the conclusion stating that BAs are better than other ML algorithms for SMP. An overall improvement in the mean MMRE values equal to 38.10%, 47.62%, 55.11%, 67.36%, 29.31%, 39.48%, and 48.81% for Abdera, Ivy, jEdit, jTDS, Log4j, Poi, and Rave datasets, respectively, has been achieved on applying BAs when compared to other ML algorithms.

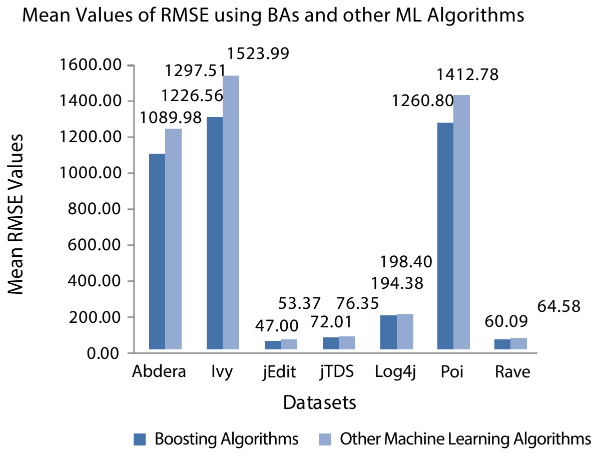


Fig. 5. Mean values of RMSE using BAs and other ML algorithms.

Subsequently, Table VIII and Table XIV present the Pred(m) values at 25%, 30%, and 75% using BAs and other ML algorithms. A comparison between these two tables also indicates BAs supremacy over other ML algorithms while predicting maintainability. Overall, comparing the best. i.e., the highest values (values marked in bold) obtained for Pred(0.25), Pred(0.30), & Pred(0.75) in both the tables, it is observed that these values are better using BAs than other ML algorithms for four out of the seven datasets, i.e., for 57.14% of the datasets (which is more than half). Further, for Poi and Rave datasets, better Pred(0.75) values have been obtained using BAs. Hence, it is clear that BAs perform better than other ML algorithms based on the Pred(m) values.

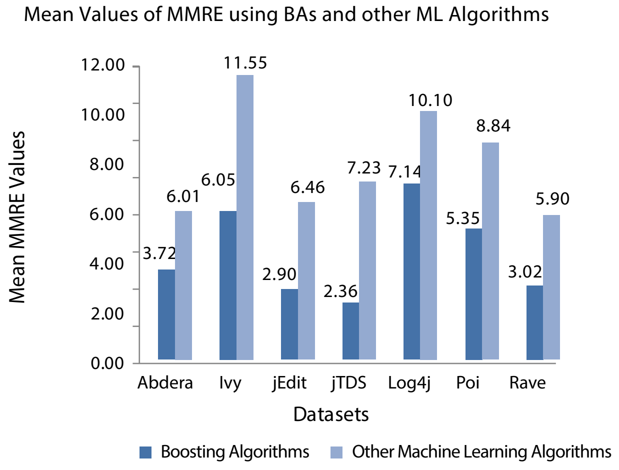


Fig. 6. Mean values of MMRE using BAs and other ML algorithms.

Therefore, on the whole, it is concluded that BAs are good performers and indeed better than other ML algorithms. All the above analysis and comparison made under this RQ, based on the RMSE, MMRE, & Pred(m) values obtained on applying BAs and other ML algorithms to each of the seven open-source datasets, further support the supremacy of BAs over other algorithms.

Conclusively, this research work can further benefit the society, especially the software engineers, in predicting the maintainability of the software being developed well in advance, thereby reducing the overall software development costs. This reduction in overall cost is mainly attributed to reducing the maintenance cost in particular, which gets accumulated with each phase of SDLC if not taken care of. The growing demand for different software in society over the last few years due to the automation of several tasks has led to a surge in the design & development of various software systems in the software industry. However, these systems require to be maintained once they are delivered to the customer involving high costs. Therefore, a great deal of specific techniques or mechanisms is needed to bring down these high costs. This can only be done by estimating the software's maintenance effort in the initial phases of development using some prediction models that can predict the software's maintainability in good time with high precision. The current research would help developers achieve this goal of predicting maintainability by utilizing different SMP models developed using various BAs, as proposed in this study. These models not only help in the task of predicting maintainability but also outperform several other models available for predictive modeling.

V. Threats to Validity

While conducting the current empirical study, certain potential threats to validity were encountered. This study has been performed on various open-source datasets, limiting its use and does not ascertain its applicability for various other types of software available in the industry for its generalization. However, a sincere attempt has been made to overcome this threat by using 10-fold cross-validation & applying all the five BAs over each of the seven datasets with different characteristics. The results obtained are possibly less biased and can further be generalized. Also, while developing prediction models using various BAs, hyper-parameter tuning of function parameters has not been performed. The default settings have been mainly used, which again becomes a limitation of this study since the results so obtained may be correct only to a first approximation. Apart from this, three of the most common threats to validity existent in any empirical study are presented below.

Internal validity refers to an extent to which conclusions of an empirical study can support the claim for cause & effect, i.e., the independent & the dependent variables. An attempt has been made to minimize this effect by applying feature selection using the RFE algorithm and using only the selected variables to study their effect on maintainability.

External validity is the extent of the generalizability of the outcomes or the results of any empirical study. A set of seven open-source datasets with different size, characteristics, and maintenance requirements has been used in this study to minimize this effect.

Construct validity is the quality of choice of various independent and dependent variables of a study, as this choice undoubtedly impacts the results of that study. So, the threat to construct validity arises from the choice of these independent and dependent variables. A set of seventeen OO metrics from different suites proposed by various researchers, namely Chidamber & Kemerer [47], Henderson-Sellers [49], and Bansiya & Davis [50] has been selected to minimize this threat rather than adhering to a particular metric suite.

VI. Conclusion & Future Direction

The current study's main objective was to analyze various ML based BAs for SMP using open-source datasets. An extensive analysis and comparison of five different BAs (AdaBoost, GBM, XGB, LightGBM, and CatBoost) were conducted using each of the seven empirically collected open-source datasets (Abdera, Ivy, jEdit, jTDS, Log4j, Poi, & Rave) to predict maintainability. Seventeen different OO metrics were selected from three different metrics suites to develop the prediction models. Feature selection using the RFE algorithm and cross-validation using the ten-fold cross-validation technique was also performed. Performance of various BAs was evaluated using RMSE, MMRE, Pred(0.25), Pred(0.30) & Pred(0.75) as the prediction accuracy measures. Further, to determine if a significant difference exists between different BAs performances & finding their mean ranks, a non-parametric statistical test named the Friedman test was conducted. Afterward, a post hoc analysis using an advanced statistical test named the Nemenyi test was also performed to identify if the difference in various BAs performance, if it exists, is statistically significant or not. Lastly, a comparison was made between the results obtained for SMP using the BAs and the results obtained on applying four other ML algorithms (DT, MLP, bagging, and EN). The major findings of the current study are as presented below.

- A reduction in features equal to 52.94% is achieved after feature selection using the RFE algorithm.
- While calculating residual errors for all the datasets using RMSE and MMRE as the accuracy measures, it was found that in the case of RMSE, GBM performed the best, followed by LightGBM, whereas, in the case of MMRE, XGB performed the best.
- Prediction accuracies also confirm the use of BAs for SMP, particularly Pred(0.75), where XGB stood out to be the best performer with a fairly reasonable predictive ability for six out of seven datasets, i.e., for 85.71% of the datasets, ranging from 51% to 79%.
- The Friedman test results and post hoc analysis using the Nemenyi test further unfolded the superiority of XGB and CatBoost BAs over other selected BAs in the study for SMP using open-source datasets.
- The comparison between the results obtained for SMP using BAs and other ML algorithms revealed that BAs are indeed the better performers than other algorithms based on all the measures of accuracy considered in this study.

Hence, prediction models developed using various BAs from the family of ML algorithms can indeed be implemented for SMP using open-source datasets. However, this is a limited implementation of the proposed study.

More research and studies can be planned in the future to implement the algorithms used in this study in isolation or in combination with other ML techniques for different types of software systems available in the industry, which are written in different programming languages to generalize the results of this study further. Different paradigms and models, more feature selection, dimensionality reduction, ensemble, and re-sampling techniques, can be considered while conducting future studies. Also, while developing prediction models in the future using the proposed algorithms, hyper-parameter tuning of different function parameters can be done as an extension to the current work.

Acknowledgment

This research work has been supported by the O/o Director (Research & Consultancy), GGSIPU under the FRGS scheme through the project entitled, "Determination of Optimum Refactoring Sequence after Prioritization of Classes on the basis of their bad smell," dt. 03.05.2019, Ref. No. GGSIPU/DRC/FRGS/ 2019/1553/62.

References

[1] "IEEE Standard for Software Maintenance," *IEEE Std 1219-1993*, 1993, doi: 10.1109/IEEESTD.1993.11557.

[2] "ISO/IEC 25010:2011(en) Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models," 2011. https://www.iso.org/obp/ui/#iso:std:iso-iec:25010:ed-1:v1:en.

[3] M. A. Ahmed and H. A. Al-Jamimi, "Machine learning approaches for predicting software maintainability: a fuzzy-based transparent model," *IET software*, vol. 7, no. 6, pp. 317–326, 2013, doi: 10.1049/iet-sen.2013.0046.

[4] N. Zighed, N. Bounour, and A.-D. Seriai, "Comparative Analysis of Object-Oriented Software Maintainability Prediction Models," *Foundations of Computing and Decision Sciences*, vol. 43, no. 4, pp. 359–374, 2018, doi: 10.1515/fcds-2018-0018.

[5] H. Alsolai and M. Roper, "Application of Ensemble Techniques in Predicting Object-Oriented Software Maintainability," in *Proceedings of the Evaluation and Assessment on Software Engineering*, 2019, pp. 370–373, doi: 10.1145/3319008.3319716.

[6] L. Kumar, D. K. Naik, and S. K. Rath, "Validating the effectiveness of object-oriented metrics for predicting maintainability," *Procedia Computer Science*, vol. 57, pp. 798–806, 2015, doi: 10.1016/j.procs.2015.07.479.

[7] R. Malhotra and A. Chug, "Application of Group Method of Data Handling model for software maintainability prediction using object oriented systems," *International Journal of System Assurance Engineering and Management*, vol. 5, pp. 165–173, 2014, doi: 10.1007/s13198-014-0227-4.

[8] H. Alsolai, "Predicting Software Maintainability in Object-Oriented Systems Using Ensemble Techniques," in *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2018, pp. 716–721, doi: 10.1109/ICSME.2018.00088.

[9] A. Chug and R. Malhotra, "Benchmarking framework for maintainability prediction of open source software using object oriented metrics," *International Journal of Innovative Computing, Information and Control*, vol. 12, no. 2, pp. 615–634, 2016.

[10] S. R. Chidamber and C. F. Kemerer, "A metrics suite for object oriented design," *IEEE Transactions on software engineering*, vol. 20, no. 6, pp. 476–493, 1994, doi: 10.1109/32.295895.

[11] R. Malhotra and A. Chug, "Software Maintainability: Systematic Literature Review and Current Trends," *International Journal of Software Engineering and Knowledge Engineering*, vol. 26, no. 8, pp. 1221–1253, 2016, doi: 10.1142/S0218194016500431.

[12] D. Michie, D. J. Spiegelhalter, C. C. Taylor, and others, "Machine learning," *Neural and Statistical Classification*, vol. 13, no. 1994, pp. 1–298, 1994.

[13] M. Sharma, S. Sharma, and G. Singh, "Performance analysis of statistical

and supervised learning techniques in stock data mining," *Data*, vol. 3, no. 4, p. 54, 2018, doi: 10.3390/data3040054.

[14] X. Zhong and D. Enke, "Predicting the daily return direction of the stock market using hybrid machine learning algorithms," *Financial Innovation*, vol. 5, no. 1, p. 4, 2019, doi: 10.1186/s40854-019-0138-0.

[15] K. C. Rasekhschaffe and R. C. Jones, "Machine learning for stock selection," *Financial Analysts Journal*, vol. 75, no. 3, pp. 70–88, 2019, doi: 10.1080/0015198X.2019.1596678.

[16] P. Kaur and M. Sharma, "Diagnosis of human psychological disorders using supervised learning and nature-inspired computing techniques: a meta-analysis," *Journal of medical systems*, vol. 43, no. 7, p. 204, 2019, doi: 10.1007/s10916-019-1341-2.

[17] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evolutionary Intelligence*, vol. 13, no. 2, pp. 185–196, 2020, doi: 10.1007/s12065-019-00327-1.

[18] M. Sharma and P. Kaur, "A Comprehensive Analysis of Nature-Inspired Meta-Heuristic Techniques for Feature Selection Problem," *Archives of Computational Methods in Engineering*, pp. 1–25, 2020, doi: 10.1007/s11831-020-09412-6.

[19] X. Ma and S. Lv, "Financial credit risk prediction in internet finance driven by machine learning," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8359–8367, 2019, doi: 10.1007/s00521-018-3963-6.

[20] H. Ghoddusi, G. G. Creamer, and N. Rafizadeh, "Machine learning in energy economics and finance: A review," *Energy Economics*, vol. 81, pp. 709–727, 2019, doi: 10.1016/j.eneco.2019.05.006.

[21] S. K. Dubey, A. Rana, and Y. Dash, "Maintainability prediction of object-oriented software system by multilayer perceptron model," *ACM SIGSOFT Software Engineering Notes*, vol. 37, no. 5, pp. 1–4, 2012, doi: 10.1145/2347696.2347703.

[22] R. Malhotra and A. Chug, "Application of evolutionary algorithms for software maintainability prediction using object-oriented metrics," in *Proceedings of the 8th International Conference on Bioinspired Information and Communications Technologies*, 2014, pp. 348–351, doi: 10.4108/icst.bict.2014.258044.

[23] L. Kumar and S. K. Rath, "Hybrid functional link artificial neural network approach for predicting maintainability of object-oriented software," *Journal of Systems and Software*, vol. 121, pp. 170–190, 2016, doi: 10.1016/j.jss.2016.01.003.

[24] M. O. Elish, H. Aljamaan, and I. Ahmad, "Three empirical studies on predicting software maintainability using ensemble methods," *Soft Computing*, vol. 19, no. 9, pp. 2511–2524, 2015, doi: 10.1007/s00500-014-1576-2.

[25] J. Zheng, "Cost-sensitive boosting neural networks for software defect prediction," *Expert Systems with Applications*, vol. 37, no. 6, pp. 4537–4543, 2010, doi: 10.1016/j.eswa.2009.12.056.

[26] E. O. Costa, G. A. de Souza, A. T. R. Pozo, and S. R. Vergilio, "Exploring genetic programming and boosting techniques to model software reliability," *IEEE Transactions on Reliability*, vol. 56, no. 3, pp. 422–434, 2007, doi: 10.1109/TR.2007.903269.

[27] M. Akour, I. Alsmadi, and I. Alazzam, "Software fault proneness prediction: a comparative study between bagging, boosting, and stacking ensemble and base learner methods," *International Journal of Data Analysis Techniques and Strategies*, vol. 9, no. 1, pp. 1–16, 2017, doi: 10.1504/IJDATS.2017.10003991.

[28] Y. Freund, "Boosting a weak learning algorithm by majority," *Information and computation*, vol. 121, no. 2, pp. 256–285, 1995, doi: 10.1006/inco.1995.1136.

[29] D. Nielsen, "Tree boosting with xgboost-why does xgboost win' every' machine learning competition?," NTNU, 2016.

[30] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[31] W. Li and S. Henry, "Object-oriented metrics that predict maintainability," *Journal of systems and software*, vol. 23, no. 2, pp. 111–122, 1993, doi: 10.1016/0164-1212(93)90077-B.

[32] M. Dagpinar and J. H. Jahnke, "Predicting maintainability with object-oriented metrics-an empirical comparison," in *10th Working Conference on Reverse Engineering, 2003. WCRE 2003. Proceedings.*, 2003, pp. 155–164, doi: 10.1109/WCRE.2003.1287246.

[33] M. M. T. Thwin and T.-S. Quah, "Application of neural networks for software quality prediction using object-oriented metrics," *Journal of systems and software*, vol. 76, no. 2, pp. 147–156, 2005, doi: 10.1016/j.jss.2004.05.001.

[34] C. Van Koten and A. R. Gray, "An application of Bayesian network for predicting object-oriented software maintainability," *Information and Software Technology*, vol. 48, no. 1, pp. 59–67, 2006, doi: 10.1016/j.infsof.2005.03.002.

[35] K. K. Aggarwal, Y. Singh, A. Kaur, and R. Malhotra, "Application of artificial neural network for predicting maintainability using object-oriented metrics," *Transactions on Engineering, Computing and Technology*, vol. 15, pp. 285–289, 2006, doi: 10.5281/zenodo.1058483.

[36] Y. Zhou and H. Leung, "Predicting object-oriented software maintainability using multivariate adaptive regression splines," *Journal of systems and software*, vol. 80, no. 8, pp. 1349–1361, 2007, doi: 10.1016/j.jss.2006.10.049.

[37] M. O. Elish and K. O. Elish, "Application of treenet in predicting object-oriented software maintainability: A comparative study," in *2009 13th European Conference on Software Maintenance and Reengineering*, 2009, pp. 69–78, doi: 10.1109/CSMR.2009.57.

[38] A. Kaur, K. Kaur, and R. Malhotra, "Soft computing approaches for prediction of software maintenance effort," *International Journal of Computer Applications*, vol. 1, no. 16, pp. 69–75, 2010, doi: 10.5120/339-515.

[39] R. Malhotra[1] and A. Chug[2], "Software Maintainability Prediction using Machine Learning Algorithms," *Software engineering : an international Journal (SeiJ)*, vol. 2, no. 2, pp. 19–36, 2012.

[40] L. Kumar and S. K. Rath, "Software maintainability prediction using hybrid neural network and fuzzy logic approach with parallel computing concept," *International Journal of System Assurance Engineering and Management*, vol. 8, no. 2, pp. 1487–1502, 2017, doi: 10.1007/s13198-017-0618-4.

[41] N. Baskar and C. Chandrasekar, "An Evolving Neuro-PSO-based Software Maintainability Prediction," *International Journal of Computer Applications*, 2018, doi: 10.5120/ijca2018916305.

[42] S. Jha *et al.*, "Deep learning approach for software maintainability metrics prediction," *Ieee Access*, vol. 7, pp. 61840–61855, 2019, doi: 10.1109/ACCESS.2019.2913349.

[43] X. Wang, A. Gegov, F. Arabikhan, Y. Chen, and Q. Hu, "Fuzzy network based framework for software maintainability prediction," *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, vol. 27, no. 5, pp. 841–862, 2019, doi: 10.1142/S0218488519500375.

[44] S. Gupta and A. Chug, "Assessing Cross-Project Technique for Software Maintainability Prediction," in *Procedia Computer Science*, 2020, vol. 167, pp. 656–665, doi: 10.1016/j.procs.2020.03.332.

[45] S. Gupta and A. Chug, "Software maintainability prediction using an enhanced random forest algorithm," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 2, pp. 441–449, 2020, doi: 10.1080/09720529.2020.1728898.

[46] S. Gupta and A. Chug, "Software maintainability prediction of open source datasets using least squares support vector machines," *Journal of Statistics and Management Systems*, vol. 23, no. 6, pp. 1011–1021, 2020, doi: 10.1080/09720510.2020.1799501.

[47] S. R. Chidamber and C. F. Kemerer, "Towards a metrics suite for object oriented design," 1991, doi: 10.1145/118014.117970.

[48] R. Malhotra and A. Chug, "An empirical study to redefine the relationship between software design metrics and maintainability in high data intensive applications," in *Proceedings of the World Congress on Engineering and Computer Science*, 2013, vol. 1.

[49] B. Henderson-Sellers, *Object-oriented metrics: measures of complexity*. Prentice-Hall, Inc., 1995.

[50] J. Bansiya and C. G. Davis, "A hierarchical model for object-oriented design quality assessment," *IEEE Transactions on software engineering*, vol. 28, no. 1, pp. 4–17, 2002, doi: 10.1109/32.979986.

[51] "MinMaxScaler Link." https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html (accessed Dec. 14, 2019).

[52] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997, doi: 10.1016/S0004-3702(97)00043-X.

[53] "RFE Documentation." https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html (accessed Dec. 16, 2019).

[54] K. T. Khaing, "Enhanced Features Ranking and Selection using Recursive Feature Elimination (RFE) and k-Nearest Neighbor Algorithms in Support Vector Machine for Intrusion Detection System," *International Journal of Network and Mobile Technologies*, vol. 1, no. 1, pp. 1832–6758, 2010.

[55] Y. Freund, R. E. Schapire, and others, "Experiments with a new boosting algorithm," in *Thirteenth International Conference on International Conference on Machine Learning (ICML'96)*, 1996, vol. 96, pp. 148–156.

[56] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997, doi: 10.1006/jcss.1997.1504.

[57] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002, doi: 10.1016/S0167-9473(01)00065-2.

[58] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[59] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[60] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 2018, pp. 6638–6648.

[61] R. Kohavi and others, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, 1995, vol. 14, no. 2, pp. 1137–1145.

[62] S. D. Conte, H. E. Dunsmore, and Y. E. Shen, *Software Engineering Metrics and Models*. Benjamin-Cummings Publishing Co., Inc. Redwood City, CA, USA, 1986.

[63] B. A. Kitchenham, S. MacDonell, L. Pickard, and M. Shepperd, "Assessing prediction systems," *The Information Science Discussion Paper Series, University of Otago*, vol. 99/14, 1999, [Online]. Available: http://hdl.handle.net/10523/1015.

[64] B. A. Kitchenham, L. M. Pickard, S. G. MacDonell, and M. J. Shepperd, "What accuracy statistics really measure," *IEE Proceedings-Software*, vol. 148, no. 3, pp. 81–85, 2001, doi: 10.1049/ip-sen:20010506.

[65] B. Iglewicz, "Robust scale estimators and confidence intervals for location," *Understanding robust and exploratory data analysis*, p. 405431, 1983.

[66] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940, doi: 10.1214/aoms/1177731944.

[67] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking classification models for software defect prediction: A proposed framework and novel findings," *IEEE Transactions on Software Engineering*, vol. 34, no. 4, pp. 485–496, 2008, doi: 10.1109/TSE.2008.35.

[68] [J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[69] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991, doi: 10.1109/21.97458.

[70] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classification," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 683–697, 1992, doi: 10.1109/72.159058.

[71] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1023/A:1018054314350.

[72] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00527.x.

Shikha Gupta

Shikha Gupta received her Master's degree in Computer Applications at Indira Gandhi National Open University (IGNOU), New Delhi, India in December 2017. She has been awarded the University Gold Medal for securing first position in order of merit and CEMCA Award 2019 for being the Best Female Student in the Master Degree Programme. Currently, she is pursuing her PhD from University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, New Delhi, India since September 2018. Her research interests include software engineering, data mining and machine learning.

Anuradha Chug

Dr. Anuradha Chug has long teaching experience of almost 30 years to her credit as faculty and in administration at various educational institutions in India. She has worked as guest faculty in Netaji Subhash Institute of Information and Technology, Dwarka, New Delhi and Regular Faculty at Government Engineering College, Bikaner. Before picking the current assignment as Assistant Professor at USICT, GGSIP University, she has also worked as Academic Head, Aptech, Meerut and Program Coordinator at Regional Centre, Indira Gandhi National Open University (IGNOU), Meerut. In academics, she has earned her doctorate degree in Software Engineering from the Delhi Technological University, Delhi, India. Before pursuing PhD, she has achieved top rank in her M.Tech (IT) degree and conferred the University Gold Medal in 2006 from Guru Gobind Singh Indraprastha University. Previously she has acquired her Master's degree in Computer Science from Banasthali Vidyapith, Rajasthan in the year 1993. Her H-index as reported by Google Scholar is 12. She has published more than 50 research papers in international and national journals and conferences. She has also served as reviewer of several national and international journals and conferences in the area of software engineering (ACM transaction, IJKESE, FOCS, IEEE Access, Neurocomputing, Informatica, Inderscience, etc). She is also the recipient of DST funded project entitled "Application of Internet of Things (IoT) in Agriculture Sector" as CO-PI. She is passionate to design, develop and deploy IoT applications in various aspects of human life. She has delivered many talks featuring Data Mining, IoT, Mining Software Repositories at various FDPs, student talks, etc.

Cross-Lingual Neural Network Speech Synthesis Based on Multiple Embeddings

Tijana V. Nosek[1]*, Siniša B. Suzić[1]*, Darko J. Pekar[2], Radovan J. Obradović[2], Milan S. Sečujski[1], Vlado D. Delić[1]

[1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad (Serbia)
[2] AlfaNum – Speech Technologies Ltd., Novi Sad (Serbia)

Received 17 December 2020 | Accepted 14 June 2021 | Published 12 November 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

The paper presents a novel architecture and method for speech synthesis in multiple languages, in voices of multiple speakers and in multiple speaking styles, even in cases when speech from a particular speaker in the target language was not present in the training data. The method is based on the application of neural network embedding to combinations of speaker and style IDs, but also to phones in particular phonetic contexts, without any prior linguistic knowledge on their phonetic properties. This enables the network not only to efficiently capture similarities and differences between speakers and speaking styles, but to establish appropriate relationships between phones belonging to different languages, and ultimately to produce synthetic speech in the voice of a certain speaker in a language that he/she has never spoken. The validity of the proposed approach has been confirmed through experiments with models trained on speech corpora of American English and Mexican Spanish. It has also been shown that the proposed approach supports the use of neural vocoders, i.e. that they are able to produce synthesized speech of good quality even in languages that they were not trained on.

Keywords

Cross-lingual, Neural Networks, Speech Synthesis, Vocoder.

DOI: 10.9781/ijimai.2021.11.005

I. Introduction

Modern text-to-speech (TTS) systems should not only be able to produce intelligible and natural-sounding speech but also to produce speech in multiple voices, styles, and preferably in multiple languages as well. Any TTS system which can handle input text in more than one language, and produce speech based on it, is referred to as multilingual TTS [1]. Multilingual TTS systems have a wide range of application. Besides being used within speech-to-speech language translation systems and interactive language tutoring systems, this functionality is necessary for a TTS system to be able to insert an occasional foreign language word into otherwise mono¬lingual speech, or even to alternate between languages in a manner consistent with the syntax and phonology of each language, which is referred to as code mixing.

The simplest solution to multilingual synthesis is based on simultaneous use of separate monolingual systems. However, since such systems are typically trained on speech corpora recorded by different speakers, this inevitably leads to inferior quality in code-mixing scenarios. On the other hand, TTS systems which can produce multiple languages in the voice of a single speaker, but typically require speech corpora from bilingual or polyglot speakers, are referred to as polyglot TTS [1], [2]. The capability to produce speech in a particular language although the training speech data in the voice of the target speaker does not contain any speech in that language is referred to as cross-lingual speech synthesis. It represents a natural alternative to using polyglot speakers for the production of training data, brought about by scientific and technological development in machine learning.

Initial attempts at producing cross-lingual synthesis were proposed for concatenative systems in the early 2000s and were based on creating phoneme mappings between source and target language [3]–[5]. Such approaches were able to generate phonetically accurate speech output, but since the intonation they were able to achieve was based on the existing source language, they were mostly applicable in code-mixing scenarios to generate some foreign words. Furthermore, sufficiently accurate phoneme mappings could be established only between languages with similar phonetic content. Another approach to concatenative cross-lingual synthesis is based on frame-level mapping. In the algorithm proposed in [6], source speaker recordings in language L1 are first spectrally warped towards target speaker recordings in language L2. Warped trajectories from source speaker are used for guiding the selection of appropriate frame-level spectral features from the target speaker, resulting in a set of utterances in L1 made from frames belonging to the target speaker. Selected target speaker frame features are then used for training a Hidden Markov Model (HMM) system capable of producing speech in the voice of the target speaker, but in L1, a language absent from the initial speech corpus of the target speaker. An extension of this approach is described in [7], proposing the use of bilinear spectral warping instead of piecewise linear, inclusion of original speech from target speaker

* Corresponding author.

E-mail addresses: tijana.nosek@uns.ac.rs (T. V. Nosek), sinisa.suzic@uns.ac.rs (Siniša B. Suzić).

in L2 into the training set for the HMM, as well as joint treatment of phonemes from both L1 and L2 based on their places and manners of articulation.

The shift of the focus in speech synthesis from concatenative to parametric approaches, brought about by the need for increased flexibility, has also influenced the development of cross-lingual speech synthesis. The first such approach, based on HMM, has enabled cross-lingual synthesis based on state mapping [8]. In this approach a bilingual speaker corpus is used to create two decision trees and an appropriate mapping between their terminal nodes is then established based on KL divergence. The obtained mapping is then applied to a monolingual speaker to generate speech in a new language. A language conversion method based on a mapping between terminal nodes of two decision trees created for average voice models is presented in [9]. A framework which attempts to factorize speaker and language features, which are modelled using a range of transforms, is presented in [10].

A major breakthrough in the development of high-quality parametric TTS did not come until the advent of neural networks. Scientific progress in this area has led to a number of different approaches to cross-lingual speech synthesis as well. For instance, in the research proposed in [11] acoustic features used to produce speech in the target language are created by a deep bidirectional long short-term memory (DB LSTM) network on the basis of phonetic posteriorgrams (PPG). The network is trained using original acoustic features of the target speaker as well as PPGs of the target speaker in the source language, obtained by a speaker-independent automatic speech recognition (ASR) system in the target language. Synthesis involves input of an arbitrary text to a general TTS in the target language (trained on any non-target speaker), which is then converted into a PPG by the ASR. The PPG features are then fed to the DB LSTM, which generates acoustic features of the target speaker in the target language, according to the input text. In [12] a deep neural network (DNN) based ASR is used to match senones from one speaker-dependent HMM TTS in the source language and another one in the target language. An example of multi-speaker and multi-language DNN TTS model is presented in [13]. This model uses separate input layers for each language and separate output layers for each speaker, while hidden layers are shared among all speakers and languages, and cross-lingual synthesis is achieved by combining corresponding input language layers and output speaker layers. In [14] unsupervised adaptation of multilingual TTS is performed by way of a search for a linguistic context which matches the available acoustic features to the greatest degree. It has been shown that a multi-speaker architecture in language L2 can be adapted by using speech data from a single speaker in language L1 to obtain TTS in language L2 in the voice of the target speaker.

Recently, end-to-end systems, which enable speech to be produced directly from text, have achieved remarkable results [15]–[17], but they require very large quantities of training data to produce synthetic speech of high quality. The end-to-end approach has also been introduced into the area of cross-lingual TTS. Most notable approaches to cross-lingual end-to-end speech synthesis based on Tacotron2 were presented in [18]–[20]. In [18] and [19] the Tacotron2 model is extended with speaker, language, tone and stress embeddings, while [18] introduces an additional adversarial speaker classifier and residual encoder. A speaker encoder based on ResCNN architecture, used for creating embeddings which condition the Tacotron2 system for predicting spectral envelopes, is presented in [20]. In all three methods shared IPA representations of phonemes are used. In [21] speaker embeddings for bilingual speakers are analysed and it has been shown that these embeddings form distinct, partly overlapping clusters. Cross-lingual speech synthesis is obtained by applying a translation of cluster embeddings learned from a bilingual speaker to a monolingual one using a Tacotron based architecture.

In spite of their great potential, a major drawback of end-to-end systems is their requirement not only for extreme computational power but also for very large quantities of speech data, which is a problem for under-resourced languages. The model that will be presented in the paper enables high quality speech synthesis, even in cross-lingual scenario, with very limited resources. It is evaluated through 5 listening tests, examining (1) whether the quality of synthesis decreases in comparison with monolingual TTS; (2) how the quality of synthesis in the original language by the proposed model compares with cross-lingual synthesis; (3) whether voice characteristics remain preserved in the cross-lingual scenario; (4) to what extent synthesis quality is degraded when the multilanguage model is adapted to a new speaker; and (5) how a neural vocoder compares to a deterministic one in the cross-lingual scenario.

The remainder of the paper is organized as follows. In Section II, we present a novel method and architecture for speech synthesis in multiple languages, in voices of multiple speakers and in multiple speaking styles, as well as speech data used in the training and evaluation of the proposed method. In Section III we give a detailed presentation of the experiments and their results. In Section IV we discuss the results of the experiments and in Section V we draw appropriate conclusions about the performance of the proposed method and outline the directions of future work.

II. Methods

This section will give a detailed presentation of the architectures of the models used in the experiments, training data, as well as specific points related to the implementation of models.

A. Models

The model proposed in this research builds upon our previous solution for monolingual speaker/style dependent speech synthesis based on embedding [22]. Both models follow the standard structure of speaker-dependent TTS, which will be outlined below.

1. Standard Speaker-Dependent TTS

A standard speaker-dependent TTS system consists of two neural networks, one for predicting phonetic segment durations, and the other for predicting acoustic features for each frame. The inputs of both networks contain linguistic information extracted from phonetically and prosodically annotated text. In order to take into account phonetic context, the inputs of both networks include not only the phonemic identity of the current phone, but the identities of phones at positions from −2 to 2 relative to the current phone. Phonemic identities are presented to the network as one-hot vectors, with some obvious exceptions, e.g. if a phone is sentence initial, features related to positions −2 and −1 are undefined and hence represented by all-zero vectors. Individual prosodic features are also presented to the network as additional inputs in binary form, and each of them represents an answer to a yes/no question typically related to the type and position of a particular prosodic event with respect to the current phone (such as: "Is the current phoneme in a stressed syllable?", "Is the number of syllables until the next phrase break greater than 3?" etc.). The acoustic network also obtains the information regarding phone durations and position of the current frame relative to its HMM state. In the synthesis phase, the outputs from the duration network are used as additional inputs to the acoustic network. The outputs of the acoustic network are typically converted to synthetic speech waveforms using an appropriate vocoder. A number of approaches have been proposed to extend such a model to enable it to handle multiple speakers and/or speech styles [22]–[24], or to adapt it to a certain speaker and/or speech style [14], [22], [25].

2. Monolingual Speaker/Style-Dependent TTS Based on Embedding

The model used as a starting point in this research, represents an extension of the standard speaker-dependent TTS, which supports multiple speakers and styles and requires a very limited quantity of speech data for adaptation [22]. This model follows the basic structure of the standard speaker-dependent TTS in that it is based on two neural networks, one predicting phone durations and the other predicting frame-level acoustic features, both using phonetic transcriptions and prosodic features as inputs.

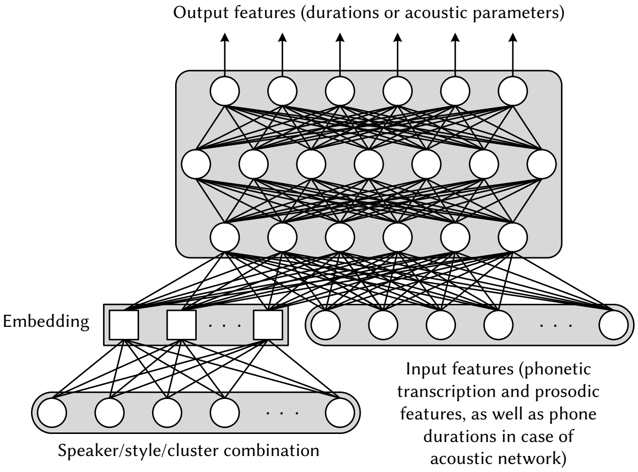

Output features (durations or acoustic parameters)
Embedding
. . .
. . .
. . .
Input features (phonetic transcription and prosodic features, as well as phone durations in case of acoustic network)
Speaker/style/cluster combination

Fig. 1. Architecture of either of the two neural networks that predict phone durations or acoustic features in monolingual speaker/style-dependent TTS based on embedding (MTTSE) [22].

The model presented in [22], capable of adaptation to different speakers and styles, uses an embedding vector as supplementary information at the input of both acoustic and duration networks, as shown in Fig. 1. This embedding vector, obtained in the training phase, uniquely represents a particular combination of speaker, speech style and cluster (a portion of the training data consistent in acoustic and prosodic quality, which typically corresponds to a single recording session). In other words, speaker ID, style ID and cluster ID are jointly represented as a single one-hot vector, which is converted into an appropriate joint embedding through training. In the resulting low-dimensional embedding space created by the network, the distance between points representing particular speaker/style/cluster (SSC) combinations reflects their similarity in terms of acoustic features and speech rate, which helps the network to efficiently generalize on unseen speech data. Hereafter, this model will be referred to as "monolingual text-to-speech based on embedding" (MTTSE).

3. The Proposed Model

The essential problem in a multilanguage scenario arises from the discrepancies between linguistic features across languages. To begin with, two languages generally do not share the same phonological inventory. Although it is usually possible to identify certain phonemes as common to multiple languages in a cross-lingual scenario, treating them as such can have negative effects since there may still be slight differences at the phonetic level. For that reason, the proposed model treats all phonemes from all languages as separate entities, which are uniquely represented as one-hot vectors, and then embedded into a low-dimensional space, as was the case with unique SSC combinations in MTTSE. The idea behind this approach is that the distance between points in the phonetic embedding space should reflect the degree of similarity between corresponding phones regardless of their language. The proposed model, hereafter referred to as "cross-lingual text-to-speech based on embedding" (CTTSE), uses 5 different embeddings for each phone in the corpus, and they are related to the phonemic identity of phones at positions from −2 to 2 relative to the current phone, as shown in Fig. 2. The size of this vector equals the sum of the sizes of phoneme inventories of all languages covered by the system, and phonetic embedding achieves efficient dimensionality reduction. Such an approach allows the network to decide e.g. to what degree the English /s/ and the Spanish /s/ are similar, and no expert knowledge is needed to match phonemes across languages.

As is the case with MTTSE, besides phonetic features, both the duration network and the acoustic network require prosodic features at their inputs. The proposed model assumes that the same prosodic annotation scheme is used for all languages included in the training. For that reason, it was possible to consider a great majority of prosodic features to be common between languages and to present them to the

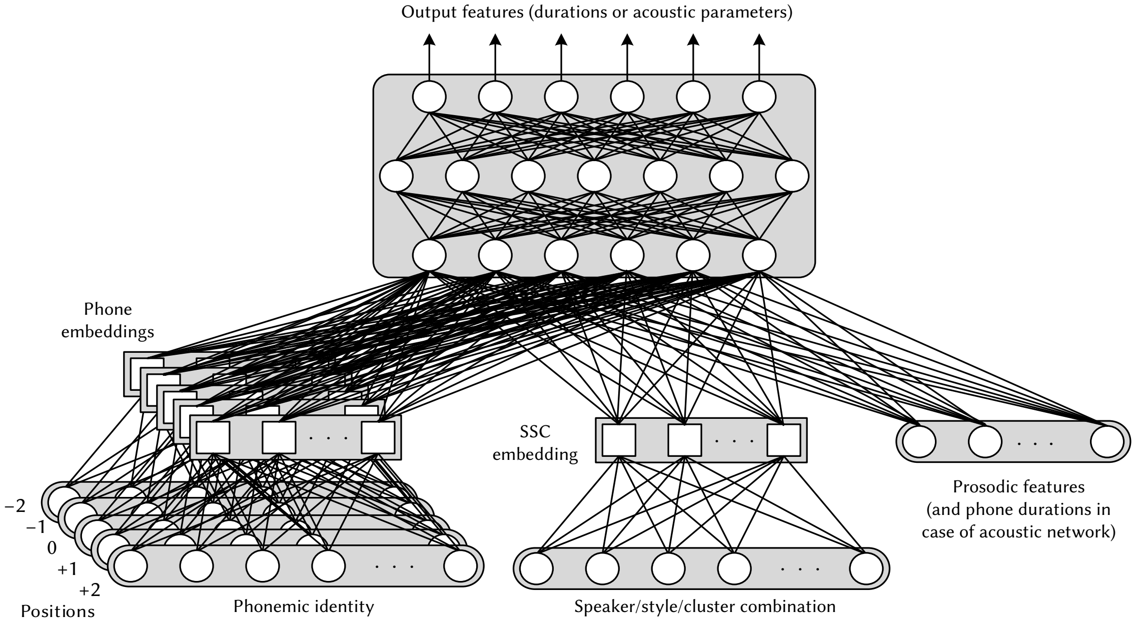

Output features (durations or acoustic parameters)
Phone embeddings
. . .
SSC embedding
. . .
. . .
Prosodic features (and phone durations in case of acoustic network)
−2
−1
0
+1
+2
Positions
. . .
Phonemic identity
. . .
Speaker/style/cluster combination

Fig. 2. Architecture of either of the two neural networks that predict phonetic segment durations or acoustic features within the model proposed for cross-lingual speaker/style-dependent TTS based on embedding (CTTSE).

neural network directly, as shown in Fig. 2. If this were not the case, prosodic inventories would also be different across languages, and some form of prosodic embedding would have been necessary.

The proposed model not only allows multi-lingual speech synthesis but cross-lingual synthesis as well. Namely, using a specific speaker-style embedding at synthesis stage will produce the voice of the desired speaker in the desired speaking style regardless of the language of the input text. Since the information of phonetic features, prosodic features and voice-related features are separated, they can be used in different combinations, enabling synthesis in the voice of a speaker who has never been "heard" speaking the target language, i.e. whose training data do not contain any speech in the target language.

Besides being able to perform cross-lingual speech synthesis, the proposed model is also capable of adaptation to the voice of a speaker not seen in the training phase, just as was the case with MTTSE, based on adaptation speech data in any of the languages in the initial model. The first phase is aimed at establishing the embedding for the new speaker/style, and it begins by random initialization of the values in the embedding layers of both networks. In this phase of the adaptation process, only the values in the embedding layers of both networks are adjusted through back-propagation while the rest of the networks is kept unchanged. The model with embedding layers adapted in such a way synthesizes speech that already resembles the target speaker/style to some extent. However, the quality of synthesized speech can be further improved through the second phase of the adaptation process, in which the same training data is used again, but the embedding layer is frozen, while the weights in the networks are modified according to the back-propagated error.

B. Data

The speech data used in this research include two languages, American English and Mexican Spanish, with a total duration of about 31 hours. All recordings were sampled at a rate of 22.05 kHz and 16 bits per sample were used. Some recordings were made in professional studios, while others were obtained from publicly available audiobooks or speeches. In the data of each speaker, styles and clusters are identified. The entire speech corpus is phonetically and prosodically annotated, and the prosodic annotation of both languages follows the Tone and Break Indices (ToBI) set of conventions, with certain extensions related to the degree of acoustic realization of pitch accents, as proposed in [26]. In other words, instead of predicting the acoustic features related to the prosody of synthetic speech based on high-level linguistic features such as part-of-speech (POS) and semantic tags, more explicit control over the prosody of synthetic speech is assumed.

The corpus of American English includes recordings of 10 female and 13 male voices, with 9 of them including speech data in more than one speaking style. On the whole, there are 82 different SSCs, ranging in length from 1.1 minutes to 1 hour, with a median of 13.8 minutes. The recordings of 17 speakers have been made in a professional studio, while the recordings of the remaining 6 speakers have been obtained from public speeches available on the Internet, and their acoustic quality is inferior. The corpus of Mexican Spanish includes recordings of 54 female and 57 male voices, however, just two of them include more than one speaking style. On the whole, there are 123 different SSCs, ranging in length from 0.8 minutes to 28.8 minutes, with a median of 1.2 minutes. Only 2 voices, the ones including multiple speaking styles, were recorded in a professional studio, while the others have been obtained from different publicly available speech corpora, and their acoustic quality is inferior. The speech rates of the two languages are also significantly different (72 ms per phone for Spanish and 117 ms per phone for English). An overview of the speech data used in this research is given in Table I.

TABLE I. Speech Data Used in the Experiment

		American English	**Mexican Spanish**
Number of speakers	Female	10	54
	Male	13	57
	Total	**23**	**111**
Speaker/ style/cluster combinations (SSC)	Total number	82	123
	Min. duration	0:01:10	0:00:50
	Max. duration	0:59:59	0:28:49
	Mean duration	0:18:58	0:02:32
	Median duration	0:13:48	0:01:13
Duration	Studio quality	22:47:39	2:48:05
	Inferior quality	3:07:04	2:24:19
	Total	**25:54:43**	**5:12:24**

C. Model Implementation

The experiments are based on two single language models (for American English and Mexican Spanish) as well as one multilanguage model. The implementation of all models used in the experiments is based on the Merlin toolkit [27] and the TensorFlow framework [28].

1. Single Language Models

The single language models used for reference in this research represent monolingual TTS systems based on embedding (MTTSE), whose architecture is described in Section II.A.2.

The duration network has an input layer of size 641 for English and 610 for Spanish, 3 feedforward layers of size 1024 with tangent hyperbolic activation functions, one LSTM layer of size 1024, and a linear output layer of size 5, for the prediction of the durations of each HMM state of a phone. The exact size of the input layer for both languages is determined by their phoneme inventories and specific implementation details, which will be illustrated in detail for the case of American English. Firstly, besides the standard 39 phonemes in the phoneme inventory, the set of phoneme identifiers used in this research also included silence and non-phonemic glottal stop. Secondly, if the articulation of a phone was significantly impaired even taking into account its context, it was labelled as "damaged" and considered as a separate phoneme (e.g. "damaged /m/" as opposed to /m/). This was taken into account only in the input section related to the phonemic identity of the current phone, while in the sections related to positions ±1 and ±2 this impairment was disregarded. Thus, the size of the input section related to the current phone is 80, while the sizes of the remaining 4 sections are 41 each (Fig. 1). Finally, 82 SSC combinations and 315 prosodic features used for English bring the total size of the input layer to 641, as mentioned previously. Following the same reasoning, the total size of the input layer for Mexican Spanish (610) can be obtained taking into account the size of its standard phonemic inventory (28), the number of existing SSC combinations (123), and the number of prosodic features used (309). It should be noted that, although essentially the same prosodic annotation scheme was used for both languages, there were nevertheless certain language-specific features of minor significance, which explains the difference between the numbers of prosodic features used for the two languages. The sizes of SSC embeddings are 10 for the English model, and 12 for the Spanish model.

The architecture of the acoustic network is basically the same, with the input layer for both languages increased by 9 to accommodate for new frame related features [27]. As in the case of the duration network, hidden layers contain 1024 neurons, while the output layer contains 130 neurons, whose outputs correspond to the values of 40 mel-generalized cepstral coefficients (MGC), 2 band aperiodicity coefficients (BAP), the value of fundamental frequency, the first and second derivatives of all features previously mentioned, as well as one feature related to the degree of voicing (VUV).

2. Multilanguage Models

The multilanguage model used in the experiments has been built along the principles of cross-lingual TTS based on embedding (CTTSE), outlined in Section II.2.

In the multilanguage model a joint phoneme inventory of size 70 is used, including two non-phonemic glottal stops (one per language) as well as silence as phoneme identifiers. As was the case with MTTSE, the current phone is represented by a one-hot vector which considers poorly articulated phones as separate phonemes. Finally, 205 SSC combinations and 286 binary prosodic features shared between the two languages bring the total size of the input layer to 908 (Fig. 2). The size of the SSC embedding is set to 15, the size of the embedding related to the current phone is set to 10, while the sizes of the embeddings related to phones at positions ±1 and ±2 are set to 5.

3. The Choice of Hyperparameters

The choice of the size of the networks was largely based on our previous research related to embedding [22]. For instance, it has been shown that a smaller network (512 neurons per hidden layer instead of 1024) would produce synthetic speech of somewhat inferior quality. While most inputs to the networks are binary (0 or 1), the 9 frame-related inputs to the acoustic network are normalized to the range [0,1] at the global level. On the other hand, the output acoustic features are standardized (mean = 0, std = 1) at the level of an individual speaker, since it has been shown that doing otherwise would greatly degrade the quality of synthetic speech in a cross-lingual scenario. In cases of both MTTSE and CTTSE, the optimizer used is stochastic gradient descent with a momentum of 0.9, and starting learning rates of 0.008 for the duration network and 0.01 for the acoustic network.

As to the choice of the sizes of particular embeddings, from a theoretical standpoint, in order to keep the volume of a hypersphere which corresponds to one SSC (or one phoneme) in the embedding space relatively constant, a logarithmic dependency between the number of SSC (or the number of phonemes) and the embedding size is implied. However, in practice the choice of the size of an embedding is complicated by a number of issues. For instance, for an SSC embedding, it is not the same if a new SSC is actually a new speaker or just a new style of an existing speaker or even just a new cluster. In the research presented in [22], varying the embedding size from 4 to 40 for 67 SSC and a single language was shown to have surprisingly little impact on the performance. In this research, the size of the SSC embedding for the single language model for English was set to 10 (the values 10, 15 and 25 have been tested). Having in mind that there are about twice as many SSCs in the Spanish data, but that they also contain many more unique speakers, the size of the SSC embedding for the single language model for Spanish was set to 12. As to the size of phoneme embeddings in the multilanguage model, they greatly depend on the actual overlap between phonemic inventories of the two languages, not only on the phonological level, but on the phonetic level as well. Since the union of phoneme inventories of English and Spanish contains 67 phonemes, and since in the case of the current phoneme a phone with impaired articulation was considered as a separate phone, the values of 5, 10 and 15 were tested as the sizes of the embedding of the phonemic identity of the current phone, and the value of 10 was chosen as the one producing the highest quality of synthetic speech. In case of phones at positions ±1 and ±2, the size of the embedding was set to 5 since they carry less important information, and the impairment in their articulation is disregarded (i.e. impaired phones are not treated as separate phonemes). It was also established that, although final synthesis does not vary much in quality, embedding space looks more sensible for specific embedding sizes. Table II illustrates the case when the embedding size is set to 10, listing the nearest neighbours for certain phonemes. It can be seen, with some exceptions, that the distance in the embedding space indeed reflects the acoustic similarity between phonemes. For instance, English and Spanish /k/ are quite similar on the phonetic level as well, which is why the network has set them closely together in the acoustic space, unlike English and Spanish /b/, whose phonetic features are somewhat different. The anomaly of English /ʔ/ being identified as the closest neighbour of Spanish /a/ remains unexplained, but it should be noted that its influence on the quality of synthesis may be minor, since the acoustic features are formed not only on the basis of the current phoneme embedding but on the basis of embeddings at positions ±1 and ±2 as well, and /ʔ/, unlike /a/, is almost exclusively found between vowels. It should also be noted that the positions of embedding of phonemes with impaired articulation in the embedding space are irrelevant since these phonemes will never be used in synthesis.

TABLE II. Nearest Neighbours of Certain Phonemes in Case the Size of the Phoneme Identity Embedding Is Set to N = 10, With Respective Euclidean Distances Given in Brackets

Phone	**1st neighbour**	**2nd neighbour**	**3rd neighbour**
Sp. /b/	En. /w/ (2.61)	Sp. /w/ (3.48)	Sp. /g/ (3.49)
Sp. /k/	En. /k/ (3.04)	Sp. /g/ (3.21)	En. /g/ (3.43)
Sp. /ɾ/	En. /r/ (4.20)	En. /ɚ/ (4.39)	En. /d/ (4.98)
Sp. /a/	En. /ʔ/ (2.49)	En. /ɑː/ (3.14)	En. /e/ (3.51)
Sp. /e/	En. /j/ (2.07)	En. /e/ (2.68)	En. /ɪ/ (2.16)
Sp. /u/	Sp. /o/ (2.41)	En./oʊ/ (3.03)	Sp. /w/ (3.41)

In both single language and multilanguage models, the duration network was trained for 100 epochs, while the acoustic network was trained for 45 epochs. Particular attention has been giving to the choice of the batch size. As the duration model is phone aligned, the batch size is represented as a product of the number of streams and the number of phonemes, where a single stream is made of concatenated sentences from the corpus. The batch size for the duration model was set to 8×50, which means that the update of weights is carried out each time a sequence of 50 phonemes from 8 different streams of sentences is processed by the network. The values given above were chosen after testing 4 to 16 streams and 16 to 50 phonemes per stream or even a single sentence as the entire batch, having in mind that for both languages the average sentence length in the corpus is close to 50 phonemes. Although it has been shown that the choice of one sentence per batch is satisfactory for synthesis of a speaker-language combination that exists in the training corpus, it is not suitable for cross-lingual scenario since it results in synthetic speech whose dynamics resemble the original language too much (e.g. an English speaker would speak Spanish too slowly). On the other hand, a batch size of 8×50 has shown to be suitable for high-quality synthesis regardless of whether the speaker-language combination exists in the training corpus or not. In the case of the acoustic network, batch size is represented as the product of the number of streams and the number of frames of length 5 ms per stream. By testing different combinations of values it has been found that, although high-quality synthesis for a speaker-language combination existing in the corpus is possible with batches as small as 32×25, cross-lingual scenario requires at least a batch size of 4×400. This implies that the update of weights should be done each time a sequence of 400 frames (corresponding to 2 seconds of speech, i.e. one half of an utterance of average length) from 4 different streams of sentences is processed by the network.

The imbalance between the representation of particular SSCs in the training corpora for both languages has been mitigated by using SSC-specific weight coefficients. Namely, when the total loss $J(\boldsymbol{\theta})_b$ for a batch is calculated, weight coefficients which boost the contributions of SSCs underrepresented in a particular training corpus are taken into account:

$$J(\boldsymbol{\theta})_b = \frac{1}{N_b}\sum_{j=1}^{N_b} w_j \sum_{i=1}^{N_{out}} \left(y_{ij} - t_{ij}\right)^2 \tag{1}$$

where N_b is the size of a batch (in samples, i.e. phones or frames), N_{out} is the size of the output layer, w_j is the weight coefficient corresponding to the SSC relevant to the j-th sample of the batch, and y_{ij} and t_{ij} are the calculated (predicted) and the target value of the i-th output for the j-th sample of the batch, respectively. The weight coefficient w_k corresponding to k-th SSC is given by:

$$w_k = \alpha\sqrt{N_k} \tag{2}$$

where N_k is the total number of utterances corresponding to k-th SSC, and α is a normalization factor given by:

$$\alpha = \sum_{k=1}^{N_{SSC}} \sqrt{N_k} \tag{3}$$

where N_{ssc} is the total number of SSCs.

4. Generation of Speech Waveforms

The first approach to the generation of speech waveforms from predicted acoustic features was based on WORLD, a widely-used deterministic vocoder [29]. It assumes a minimum phase for the spectrum and by using the predicted acoustic features it converts the cepstral features into a linear amplitude spectrum and produces excitation signal by mixing a pulse and a noise signal in the frequency domain, where each frequency band is weighted by a value of predicted band aperiodicity acoustic features. Finally, it generates a speech waveform based on the source-filter model. The second approach was based on WaveRNN [30], an increasingly popular neural vocoder, which predicts the more significant and the less significant halves of the 16-bit output sample separately, and supports simultaneous prediction of several output samples. Since it requires extreme processing power and large quantities of training data per speaker, it has been tested just for a single speaker who was most represented in the available training data, in order to establish whether the cross-lingual scenario is possible with a neural vocoder and how the use of a neural vocoder instead of a conventional one affects the quality of cross-lingual synthesis.

III. Experiments and Results

A. Experiment 1: Single Language Vs. Multilanguage Model

The aim of Experiment 1 is to compare the quality of speech generated using the multilanguage (ML) model and the single language (SL) model in a speaker's native language. Although ML in this experiment supports only two languages, it can be easily extended to more languages. It should be emphasized that in this research an extreme disbalance between the corpora of two languages exists – the English corpus is 5 times bigger than the Spanish one (~25h vs ~5h) but includes far fewer different speakers (23 vs 111, or 83 unique SSC vs 124 unique SSC).

Since there are original and synthesized recordings of the same sentences (withheld during training) it was possible to conduct objective evaluation of the quality of synthesized speech, based on a comparison of values of acoustic features extracted from original recordings and those predicted by the TTS model. The standard measures are: mel-cepstral distance (MCD), root mean square error of the fundamental frequency (RMSE F0), correlation between predicted and true fundamental frequency (CORR F0), root mean square error of phoneme duration expressed in frames per phone (RMSE DUR) as well as correlation between predicted and true phoneme durations (CORR DUR). Table III shows the objective measures for each language.

Since subjective evaluation is still considered in the literature as the most reliable way of establishing the quality of speech synthesis, a comparison between synthetic speech obtained by ML and SL models was also carried out through a preference test including 31 non-native listeners, who declared themselves as having sound knowledge of both English and Spanish. Each listener was given 20 tasks (10 per language). In each task there were two sentences with the same linguistic content – one sentence synthesized by SL model and the other one by ML model. In the preference test each language was represented by 5 speakers, 2 male and 3 female ones. The listeners were asked to select the utterance of better quality in terms of intelligibility and naturalness, and the answer "no preference" was also acceptable. The results of the preference test are given in Fig. 3.

TABLE III. Objective Measures of Distance Between Synthesized and Natural Speech

		MCD (dB)	RMSE F0 (Hz)	CORR F0	RMSE DUR	CORR DUR
English	SL	5.26	32.30	0.90	5.79	0.84
	ML	5.39	33.34	0.89	5.58	0.85
Spanish	SL	5.29	24.39	0.91	5.68	0.77
	ML	5.19	24.39	0.91	5.61	0.78

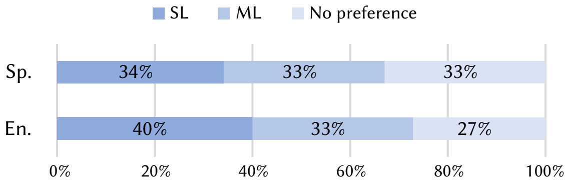


Fig. 3. Results of subjective comparison of the quality of speech synthesized by SL and ML.

B. Experiment 2: Speech Quality in a Cross-Lingual Scenario

The aim of Experiment 2 was to evaluate the quality of cross-lingual speech synthesis. Since in a cross-lingual scenario, ground truth examples (original recordings) do not exist, the only way of testing the quality of synthesis is subjective evaluation. Since it included rating the utterances delivered by different target speakers on a MOS scale rather than simple comparison, only native speakers of English and Spanish participated in the listening tests.

Two listening tests were carried out with 2 groups of 21 listeners per group – one for English and the other for Spanish. In each test there were 10 tasks, containing 4 utterances each. In each task the content of all utterances was the same, but two were synthesized by a speaker-language combination that exists in the training corpus, while the other two were synthesized by a speaker-language combination that does not exist in the corpus (i.e. cross-lingual scenario). Each of the 4 utterances in a task was delivered in the voice of a different speaker, of whom 2 were native English (male and female) and 2 native Spanish (male and female). In the entire test containing 10 tasks, there are sentences from 8 different speakers. In each task, listeners were asked to evaluate the quality of 4 synthesized sentences in terms of intelligibility and naturalness on a 1 to 5 MOS scale. Multiple speakers were introduced to neutralize any bias that a listener may have towards a specific voice. The results of the experiment are presented in Fig. 4. and Fig. 5.

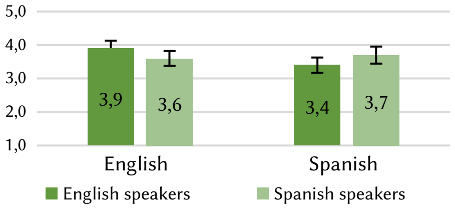


Fig. 4. Results of subjective comparison of the quality of original-language and cross-lingual synthesis (mean values with 95% confidence intervals are shown).

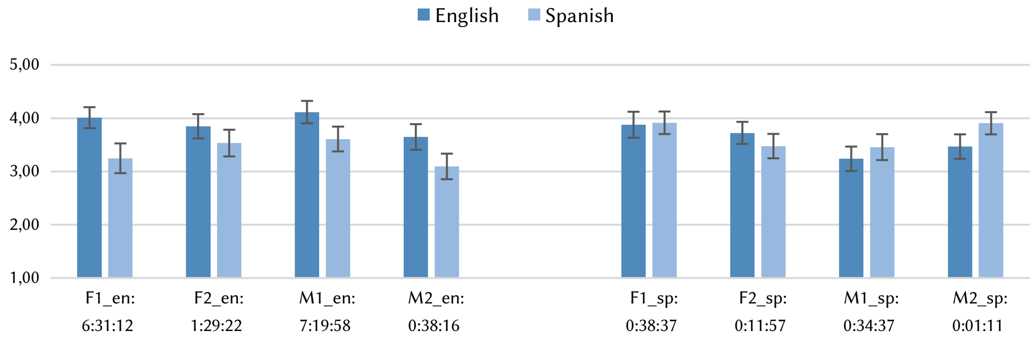


Fig. 5. Results of subjective comparison of the quality of original-language and cross-lingual synthesis for individual speakers (suffixes 'en' and 'sp' indicate the original language of each speaker). The amounts of available training data are also indicated.

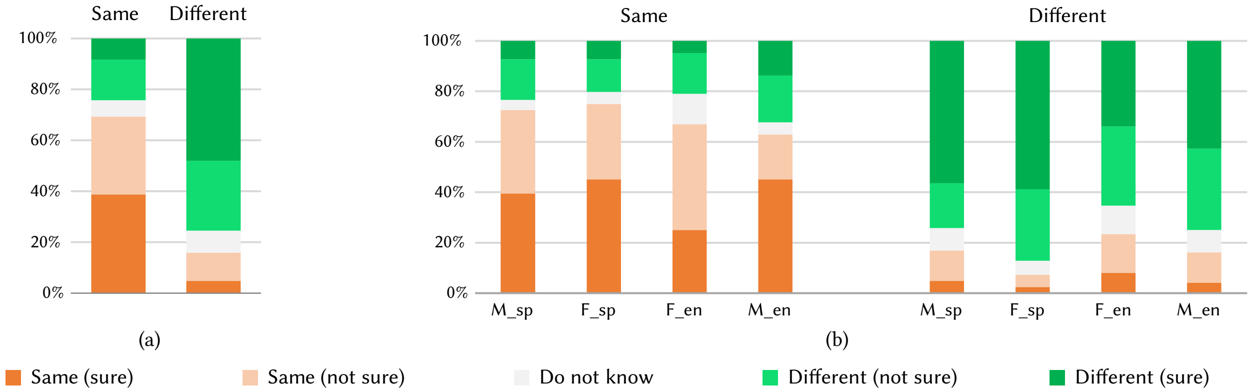


Fig. 6. Results of the evaluation of voice similarity in case of cross-lingual speech synthesis: (a) overall; (b) for each target speaker individually. Labels 'Same' and 'Different' indicate whether both sentences in a pair were delivered by the same speaker.

C. Experiment 3: Voice Similarity in a Cross-Lingual Scenario

The aim of Experiment 3 was to establish to what extent the characteristics of a speaker's voice remain preserved in the cross-lingual scenario. However, the task of evaluating voice characteristics is not easy when the entire sentence, even the language, is different. For that reason, the test has created as follows.

Each of the 32 tasks in a test included a pair of utterances and the listeners were asked to state their opinion as to whether both utterances were delivered by the same "virtual speaker" on a 1 to 5 scale defined as follows:

1. I am sure the utterances were delivered by different speakers;
2. I think they were delivered by different speakers;
3. I do not know whether they were delivered by the same speaker;
4. I think they were delivered by the same speaker;
5. I am sure they were delivered by the same speaker.

In each pair, one utterance was in Spanish and the other in English, both produced by the ML model. Consequently, in case of pairs of sentences from the same speaker, one sentence is necessarily synthesized using the cross-lingual scenario. There were 8 tasks for each of the 4 different target speakers (one for each combination of gender and native language). Half of the pairs of sentences in 32 tasks were delivered by the same speaker, while in the other half the utterances were delivered by different, but similar speakers. The test was presented to 31 non-native listeners, and the results are shown in Fig. 6.

D. Experiment 4: Adaptation to a New Speaker

The aim of Experiment 4 is to establish whether it is necessary to retrain the entire multispeaker (MS) ML model when a new speaker appears in order to obtain cross-lingual synthesis, or it is sufficient to adapt the existing MS ML model to new speaker data, as described in Section II.A.3. In the experiment two new native English speakers were introduced – a female, whose training corpus can be considered as small (10 min, 4 SSCs, inferior quality), and a male, whose training corpus can be considered as being of moderate size (45 min, 3 SSCs, studio quality).

The preference test consisted of 20 pairs of utterances, 10 per speaker, of which 5 were in English (original speaker-language combination) and 5 in Spanish (cross-lingual scenario). In each pair of utterances, one was synthesized by the multispeaker model that included the target speaker in the training corpus, while the other was synthesized by the model which had been adapted to the target speaker. A total of 31 non-native listeners were asked to select the utterance of better quality in terms of intelligibility and naturalness, and the answer "no preference" was also acceptable. The results are shown in Fig. 7.

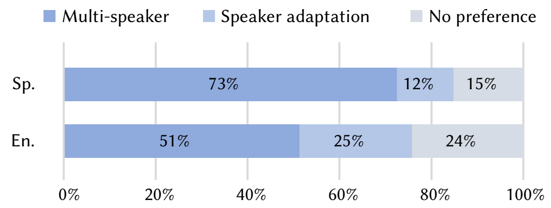


Fig. 7. Results of subjective comparison of the quality of multi speaker and speaker adaptation synthesis.

E. Experiment 5: Deterministic Vs. Neural Vocoder

The aim of Experiment 5 is to compare results that can be obtained by using a deterministic vocoder and from the data-driven vocoder, both in original-language and cross-lingual scenario. In this research

WORLD vocoder was used as a widely-used example of a deterministic vocoder, while WaveRNN was selected as an increasingly popular and efficient data-driven vocoder. At first sight, the experiment may not seem immediately related to this research since any vocoder produces speech samples given appropriate acoustic features at its input, and thus it may not be obvious that the purpose of the experiment is anything beyond a simple comparison between WORLD and WaveRNN. However, it should be taken into account that a neural vocoder has to be trained in order to produce speech based on acoustic features, and in the cross-lingual scenario, training data for a specific voice in the target language does not exist. In other words, the vocoder is required to produce speech in a language it has not heard before, and the experiment aims to establish whether this is at all practicable, and whether the possible loss in the quality of synthesized speech is acceptable.

A comparison between synthetic speech obtained by WORLD and WaveRNN was carried out through a preference test including 31 non-native listeners, who declared themselves as having sound knowledge of both English and Spanish. Each listener was given 10 tasks (5 per language). In each task the listeners were presented with two sentences with the same linguistic content – one sentence synthesized using WORLD and the other one using WaveRNN. All speech samples used in the experiment were synthesized in the voice of the English speaker (a female one, neutral style) with the most data in the training corpus. The samples representing original speaker-language combinations and cross-lingual synthesis were equally represented and randomly distributed in the test. The listeners were asked to select the utterance of better quality in terms of intelligibility and naturalness, and the answer “no preference” was also acceptable. The results of Experiment 5 are presented in Fig 8.

IV. Discussion

In this section an interpretation of the results will be presented, outlining their importance in view of the recognized limitations of the study. To illustrate the quality of speech synthesized by different approaches described in the paper, and to substantiate the results of the listening tests, the speech samples used in the tests have been made available at: www.alfanum.ftn.uns.ac.rs/crosslingual.

A. Experiment 1: Single Language Vs. Multilanguage Model

Based on objective measures of distance between acoustic features of original and synthesized speech (Table III), it can be concluded that in most cases the performance in a certain language of the ML model matches the performance of SL models. The results related to phone durations are slightly improved, as well as acoustic measures for the language with less data. This is in line with the expectation that, given enough training data, a SL model is capable of producing synthetic speech of high quality, and the use of a ML model may be advantageous since it is capable of overcoming the lack of training data for underrepresented languages. On the other hand, the results of the subjective evaluation (Fig. 3) show that there is no significant difference between compared methods, although the SL model is slightly preferable than the ML model in the case of English. This can be explained by the fact that in case of English much more training data were available, and since the SL model had enough material for training, adding a new language was only a distraction for the network. On the other hand, in case of Spanish, a three-way tie among the preference of SL, preference of ML and no preference at all, indicates that the additional language was not a distraction, although it did not improve the Spanish synthesis either. Our expectation that the benefit from the use of ML model increases with the increase of the number of underrepresented languages as well as the scarcity of training data, will be the subject of further research, as soon as prosodically annotated data in more languages become available.

It should be noted that a similar analysis has been conducted in [18], including a comparison of the quality of synthesis from end-to-end SL and ML models. Speech data used in [18] also exhibited a significant imbalance regarding the representation of each language (387 hours of English vs. 97 hours of Spanish and 68 hours of Chinese). The ML model was based on three languages – English, Spanish and Chinese. It was concluded that there was no significant degradation in case of ML, although SL was found to be slightly preferable (MOS grades were 0.1 lower for ML on average). It was also confirmed that SL was especially preferable in case of English (the difference in grades was more than 0.2), than in case of other two languages (the difference in grades was less than 0.1).

B. Experiment 2: Speech Quality in a Cross-Lingual Scenario

From the results of Experiment 2, shown in Fig. 4 and Fig. 5, it can be seen that, in general, higher quality synthesis is achieved for English, which could be expected since the English training corpus is 5 times larger than the Spanish one. However, taking into account only original speaker-language combinations (i.e. cases where speech is synthesized in the original language of the speaker), it is interesting to note that synthesis of Spanish speech was, on average, rated only 0.2 lower than synthesis of English speech. This is a very encouraging result, having in mind the difference in the sizes of training corpora. More importantly, the results of Experiment 2 fully confirm the feasibility of cross-lingual speech synthesis based on CTTSE. Namely, the quality of cross-lingual synthesis has been rated as inferior to original-language synthesis by only 0.3 in both cases (English to Spanish and vice versa).

Analysing grades obtained for each individual speaker, it can be noticed that all target English speakers have grades higher than 3.5 for synthesis in English, 2 of them even grades higher than 4.0, while the target Spanish speakers have obtained lower grades for synthesis in Spanish, 2 of them even less than 3.5. This may be explained by the fact that all English speakers have much larger training corpora than their Spanish counterparts. However, it is interesting to note that the Spanish speaker with only 1 minute of training data and the English speaker with more than 7 hours of training data obtained grades which differ by only 0.21 for their native languages. The differences between synthesis in the original language and cross-lingual scenario are smaller (up to 0.45) in case of Spanish speakers, where in one case the cross-lingual scenario was graded even better than synthesis in the original language. For English speakers, the cross-lingual scenario is graded usually as at least half a grade inferior to synthesis in the original language. However, no speaker-language combination obtained a grade below 3.1.

Although it was not mentioned in the comments of listeners since they were unaware of the origin of each utterance, it can safely be concluded that inadequate speech rate is one of the factors that have reduced the quality of cross-lingual speech synthesis. By comparing cross-lingual speech samples, it can be concluded that synthesized voices tend to preserve the speech rate of their original language, which implies that English voices speak Spanish unnaturally slow, while Spanish voices speak English unnaturally fast. This may be the consequence of the specific approach to network output normalization, which will be investigated further.

The quality of synthesis for underrepresented languages in [18] was graded as equal or slightly inferior to the case of the language which was represented with the most training data for an original speaker-language combination (e.g. for Chinese the average MOS grade was approximately 0.3 lower than for English and Spanish). Switching to a cross-lingual scenario introduced a slight degradation in quality

(0.06 lower MOS grade in case of 3 languages, and 0.13 in case of only Spanish to English and vice versa). Although these results are better than in our research, it should be noted that in our research the amount of training data was 15 to 20 times smaller. Another research, presented in [13], contrary to our research and [18], used corpora of bilingual speakers in order to construct a ML model. They also conducted an experiment with 2 bilingual speakers and 1 monolingual speaker and tested the cross-lingual scenario (about 45 minutes of data for each speaker-language combination was used). The quality of synthesis obtained by the ML model in a cross-lingual scenario was graded with a MOS grade by 0.25 lower than in case of a SL model trained on data from only one speaker (standard TTS).

C. Experiment 3: Voice Similarity in a Cross-Lingual Scenario

Experiment 3 aimed at establishing to what degree speech synthesized in another language retains the voice characteristics of the original speaker. As explained in Section III.C, the participants were asked to state whether they believe that each of the two utterances presented in a pair was delivered by the same "virtual speaker" on a 1 to 5 scale. If grades 5 and 4 can be considered as correct answers and grades 1 and 2 as wrong answers for pairs where the utterances correspond to the same speaker, and vice versa if they correspond to different speakers, the listeners answered correctly in 72% cases, could not decide in 8%, and gave the wrong answer in 20%. Fig. 6. provides a more detailed analysis of the results. Since the listeners recognized correctly that the speaker was the same in almost 70% cases, being sure in their answers (grade 5) in almost 40% cases, it can be concluded that the voice characteristics remained preserved in cross-lingual scenario. On the other hand, in case when the sentences in the pairs were actually delivered by different speakers, listeners correctly recognized it in almost 80% of cases, being sure (grade 1) in almost 50%. It can be noted that for the female English speaker the listeners were less sure in their answers and also the most indecisive in pairs where her voice was present.

It should be noted that the reported degradation in voice similarity in [18] in comparison with the original speaker-language case was as high as 1.0. The authors of [18] have also emphasized the problem of grading voice similarity in case the sentence or even the language is different, which is why we have opted for a different approach – to ask the listeners to identify whether the two utterances in different languages have been delivered by the same speaker. On the other hand, the evaluation of the voice similarity in [13] was quite simple. Namely, owing to the use of bilingual speakers, it was possible to directly compare the result of synthesis from the ML model in a cross-lingual scenario with an original recording of the speaker in the target language. A decrease of the MOS grade by 0.58 with respect to the synthesis from SL model was reported.

D. Experiment 4: Adaptation to a New Speaker

The results of Experiment 4, shown in Fig. 7, indicate that re-training the entire MS ML model from scratch including the new speaker produces speech of better quality than speaker adaptation (SA). The preference of MS ML over the SA approach is more emphasized in the case of Spanish, i.e. in case of the cross-lingual scenario. It can be assumed that, in adapting the existing model to the new speaker, the network is less ready to generalize and produce a new speaker-language combination because it overfits to the single speaker-language combination used for adaptation.

The results do not differ much depending on whether the training corpus is small or of moderate size, although SA has shown to be more acceptable in the case of the speaker with a moderate training corpus. It is interesting to note that during speaker adaptation, embedding values for the phonemes of the language not existing in the corpus of the new speaker will not be updated. However, this should not lead to a significant difference in quality with respect to the cross-lingual scenario in which a speaker is included in the training of the original ML model, since in that case his/her data will influence only the embeddings for the phonemes of languages that exist in his/her training corpus.

E. Experiment 5: Deterministic Vs. Neural Vocoder

From the results of Experiment 5, shown in Fig. 8, it can be seen that, while WaveRNN is preferable over WORLD in both English and Spanish, i.e. in both original language and cross-lingual synthesis, the preference in case of English is negligible. As is well known, both vocoders have their own specific properties, e.g. while synthesis by WORLD is relatively stable but with a constant impairment in quality referred to as "buzzing" [31], the synthesis by WaveRNN generally sounds more natural but is less stable. A point of some relevance for this research is that WaveRNN synthesis includes a certain overtone which may affect the timbre of the voice, but it could not be spotted by listeners to whom the original voice is unknown. It should also be noted that, unlike WORLD, WaveRNN exhibits significant flexibility in terms of architecture and hyperparameters, so the results can be further improved. However, a downside of WaveRNN is the necessity of large corpora, and in this experiment, only one speaker with the sizable corpus was used, so its adaptation or multispeaker training are the subjects of further research.

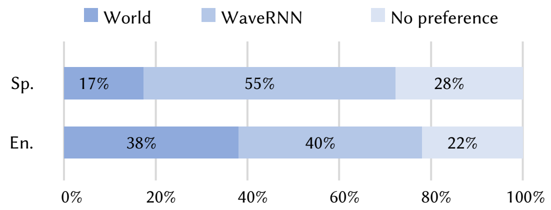


Fig. 8. Results of subjective comparison of the quality of utterances obtained by using WORLD and WaveRNN vocoders.

Most importantly, the experiment has shown that WaveRNN is able to produce high-quality synthetic speech even in the language it is not initially trained on, which confirms the assumption that acoustic features from the TTS model of one speaker are meaningful input to WaveRNN regardless of the language to which they may correspond.

V. Conclusion

The study presents a novel method for multilingual and cross-lingual neural network speech synthesis. Firstly, it shows that the proposed method is capable of speech synthesis in multiple languages, and that it is a good basis for the creation of speech synthesis for languages in which a relatively small quantity of speech data is available. As its main point, the study shows that it is possible to synthesize speech in a specific person's voice in a language that this person has never spoken. The quality of cross-lingual synthesis in terms of intelligibility and naturalness, as well as the resemblance of the synthesized voice in a cross-lingual scenario to the same voice in original language synthesis, were both established to be relatively high (a difference in quality on a MOS scale was found to be 0.3). Since it would be impractical to retrain the entire system each time a new speaker is introduced, a method for speaker adaptation in a cross-lingual scenario was examined as an alternative and it was found that does not lead to an unacceptable loss in speech quality, particularly in the case of the language with greater overall quantity of training data. Finally, it has been shown that the proposed method for cross-lingual synthesis supports the use of neural vocoders, even though it means that they have to be trained on data in one language, and used for synthesis of speech in another. The study, thus, brings the state of the art in speech technology one step closer to the synthesis of arbitrary text in an arbitrary voice, speaking style and language, easily extensible to new speakers, styles and languages.

The study is somewhat limited by the fact that it was based on only two languages, with significant differences in both the number of speakers in the training corpus as well as the average quantity of available data per speaker. However, most of its results and conclusions are in agreement with expectations based on theoretical knowledge. Our future research on this topic will include the extension of the study to multiple languages as soon as more data become available. The study also raises a number of questions related to specific implementation of particular models, most notably the normalization of network outputs, which will also be investigated further in our future work.

Acknowledgment

This research was supported by Speech Morphing Systems Inc., Campbell, CA, United States of America, as well as the Science Fund of the Republic of Serbia (grant #6524560, AI – S-ADAPT). Speech corpora used in the research were provided by Speech Morphing Systems Inc. for research purposes.

References

[1] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," in *Proceedings of the 6th European Conference on Speech Communication and Technology EUROSPEECH 1999*, Budapest, Hungary, 1999, pp. 835–838.

[2] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft Mulan – a bilingual TTS system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2003*, Hong Kong, China, 2003, vol. I, pp. 264–267.

[3] N. Campbell, "Foreign-language speech synthesis," in *Proceedings of the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis,* Jenolan Caves, Australia, 1998, pp. 177–180.

[4] M. Moberg, K. Pärssinen, and J. Iso Sipilä, "Cross-lingual phoneme mapping for multilingual synthesis systems," in *Proceedings of the 8th International Conference on Spoken Language Processing ICSLP 2004,* Jeju Island, Korea, 2004, pp. 1029–1032.

[5] L. Badino, C. Barolo, and S. Quazza, "Language independent phoneme mapping for foreign TTS," in *Proceedings of the 5th ISCA Workshop on Speech Synthesis*, Pittsburgh, PA, USA, 2004, pp. 217–218.

[6] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2011*, Prague, Czech Republic, pp. 5120-5123.

[7] J. He, Y. Qian, and F. K. Soong, "Turning a monolingual speaker into multilingual for a mixed-language TTS," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association INTERSPEECH 2012*, Portland, OR, USA, 2012, pp. 963–966.

[8] Y. Qian, H. Liang and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin–English) TTS," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009, doi: 10.1109/TASL.2009.2015708.

[9] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association INTERSPEECH 2009*, Brighton, United Kingdom, 2009, pp. 528–531.

[10] H. Zen et al., "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1713–1724, 2012, doi: 10.1109/TASL.2012.2187195.

[11] L. Sun, H. Wang, S. Kang, K. Li, and H. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association INTERSPEECH 2016,* San Francisco, CA, USA, 2016, pp. 322–326.

[12] F. L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN approach to cross-lingual TTS," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2016*, Shanghai, China, 2016, pp. 5515–5519.

[13] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2016*, Shanghai, China, 2016, pp. 5540–5544.

[14] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Unsupervised speaker adaptation for DNN-based TTS synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2016*, Shanghai, China, 2016, pp. 5135–5139.

[15] J. Sotelo et al., "Char2wav: End-to-end speech synthesis," in *Proceedings of the 5th International Conference on Learning Representations ICLR 2017,* Toulon, France, 2017, pp. 1–6.

[16] Y. Wang et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," arXiv preprint arXiv:1703.10135, 2017. Accessed: July 15, 2020. [Online]. Available: https://arxiv.org/abs/1703.10135.

[17] S.Ö. Arık et al., "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning PMLR*, Sydney, Australia, 2017, vol. 70, pp. 195–204.

[18] Y. Zhang et al., "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," arXiv preprint arXiv:1907.04448, 2019. Accessed: July 15, 2020. [Online]. Available: https://arxiv.org/abs/1907.04448.

[19] M. Chen et al., "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding", in *Proceedings of the 20th Annual Conference of the International Speech Communication Association INTERSPEECH 2019,* Graz, Austria, 2019, pp. 2105–2109.

[20] Z. Liu and B. Mak, "Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers," arXiv preprint arXiv:1911.11601, 2019. Accessed: July 15, 2020. [Online]. Available: https://arxiv.org/abs/1911.11601.

[21] S. Maiti, E. Marchi, and A. Conkie, "Generating multilingual voices using speaker space translation based on bilingual speaker data," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2020*, Shanghai, China, 2020, 7624–7628.

[22] M. Sečujski, D. Pekar, S. Suzić, A. Smirnov, and T. Nosek, "Speaker/style-dependent neural network speech synthesis based on speaker/style embedding", *Journal of Universal Computer Science*, vol. 26, no. 4, pp. 434–453, 2020.

[23] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association INTERSPEECH 2016,* San Francisco, CA, USA, 2016, pp. 2278–2282.

[24] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2015*, South Brisbane, Australia, pp. 4475-4479.

[25] T. Delić, S. Suzić, M. Sečujski, and D. Pekar, "Rapid development of new TTS voices by neural network adaptation," in *Proceedings of the 17th International Symposium INFOTEH-JAHORINA,* Jahorina, Bosnia and Herzegovina, 2018, pp. 1–6.

[26] M. Sečujski, S. Suzić, S. Ostrogonac, and D. Pekar, "Learning prosodic stress from data in neural network based text-to-speech synthesis," *SPIIRAS Proceedings*, vol. 4, no. 59, pp. 192–215, 2018, doi: 10.15622/sp.59.8

[27] Z. Wu, O. Watts, and S. King, "Merlin: an open source neural network speech synthesis system", in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, 2016, pp. 218–223.

[28] M. Abadi et al., "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation OSDI 2016*, Savannah, GA, USA, 2016, pp. 265-283.

[29] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp.1877–1884, 2016, doi: 10.1587/transinf.2015EDP7457.

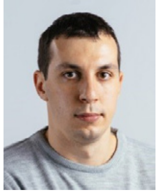
[30] N. Kalchbrenner et al., "Efficient neural audio synthesis," arXiv preprint

arXiv:1802.08435, 2018. Accessed: July 18, 2020. [Online]. Available: https://arxiv.org/abs/1802.08435.

[31] S. King, "An Introduction to Statistical Parametric Speech Synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

Tijana Nosek

Tijana Nosek, MSc, born in 1992, is a doctoral candidate at the University of Novi Sad, Faculty of Technical Sciences, Department for Power, Electronics and Telecommunication Engineering, Chair of Telecommunications and Signal Processing. She is engaged as teaching assistant at the Faculty of Technical Sciences, at courses principally related to machine learning and digital signal processing. She has participated in a number of national and international scientific projects and currently is a member of the team which participates in the scientific project Speaker/Style Adaptation for Digital Voice Assistants Based on Image Processing Methods (S-ADAPT), financed by the Science Fund of the Republic of Serbia, with the aim of improving machine learning algorithms for speech synthesis with limited training data. She also cooperates closely with "AlfaNum – Speech Technologies", a leading company in speech technologies in the region. The main area of her research is speech synthesis based on neural networks with particular focus on expressive speech. She authored or co-authored 4 research papers in renowned international journals as well as 19 papers at scientific conferences related to speech processing. She is a member of international organizations such as IEEE and Audio Engineering Society (AES).



Siniša Suzić

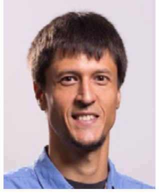
Siniša Suzić, PhD, was born in 1988. He defended his bachelor, master and PhD thesis at the University of Novi Sad, Faculty of Technical Sciences, in 2011, 2012 and 2019, respectively. He is currently engaged as teaching assistant at the University of Novi Sad, Faculty of Technical Sciences, Department for Power, Electronics and Telecommunication Engineering, at several courses related to acoustics and signal processing. He has participated in a number of national and international scientific projects related to speech processing and human-machine interaction. His areas of scientific interest include expressive speech synthesis as well as speech synthesis with limited training data. He also contributed to the creation of speech resources used in the development of large vocabulary speech recognition for Serbian and other South Slavic languages, as well as speech recognition for mobile phones. He also contributed to the development of different commercial, speech related software in Serbian and other South Slavic languages. He authored or co-authored 4 research papers in renowned international journals as well as more than 25 papers at scientific conferences related to speech processing. He is a member of IEEE.

Darko Pekar

Darko Pekar, MSc, was born in 1972. He graduated in 1998, at the University of Novi Sad, Faculty of Technical Sciences. Up to 2003 he was the leading expert at a university research project where he obtained wide-ranging experience in the field of speech technology, as well as management of scientific and technological projects. In 2003 he became CEO of the company "AlfaNum", and in cooperation with University of Novi Sad, he has continued to manage teams working on speech technology and machine learning. He currently leads the team of 25 software engineers and accompanying staff. His major achievements over the years include the development of: high quality text-to-speech systems for Serbian, Croatian, English, Spanish and Hebrew languages; large vocabulary speech recognition systems for Serbian and Croatian languages; methods for speaker adaptation by using very small quantities of speech data. Until 2019 he was also engaged as research assistant at the Faculty of Technical Sciences, and is currently finishing his PhD thesis in the area of speaker adaptation. Although he focuses on practical development and providing market-ready solutions, he has also published more than 100 articles and papers in national and international scientific journals.

Radovan Obradović

Radovan Obradović, MSc, was born in 1969 and in 1999 he received his BSc/MSc degree in Electrical Engineering from the University of Novi Sad, Faculty of Technical Sciences. He has worked in industry as a researcher in the fields of digital signal processing, speech recognition and synthesis, natural language processing and computer vision. During his cooperation with the company "AlfaNum", he has played a major role in the development of a range of speech technology solutions for Serbian and other languages, related to both speech recognition and synthesis. His current research interest includes neural speech synthesis, speech recognition, dialogue systems, applications of sparse representations in artificial neural networks, biologically inspired learning and meta learning.



Milan Sečujski

Milan Sečujski, PhD, was born in 1975, and currently works as Associate Professor at the University of Novi Sad, Faculty of Technical Sciences, Department for Power, Electronics and Telecommunication Engineering, Chair of Telecommunications and Signal Processing. He is engaged as a lecturer in university courses related to digital signal processing, time series analysis as well as machine learning. His areas of scientific interest include computational linguistics, speech and language technology, as well as human-machine interaction. The result of his master research thesis evolved into the highest quality speech synthesizer in Serbian. This system has since been widely used, initially by the blind and visually impaired computer users, but its use has since spread into the domain of telecommunications services, where it has remained the most widely used speech system in Serbian and Croatian. He has participated in a number of international projects, and is currently participating in the Erasmus+ project BENEFIT (Boosting the Telecommunications Engineer Profile to Meet Modern Society and Industry Needs). His current research includes natural language processing as well as mathematical modeling of the prosodic features of speech, most notably for the purposes of expressive speech synthesis as well as speaker conversion in speech synthesis. Apart from his work in the domain of speech technology, he has made scientific contribution in the field of acoustic metamaterials as well. He is a member of international organizations such as IEEE and Audio Engineering Society (AES).

Vlado Delić

Vlado Delić, PhD, was born in 1964. He is engaged as Full Professor at the University of Novi Sad, Faculty of Technical Sciences, Serbia, and he is also the head of the Chair of Communication Engineering and Signal Processing. He received the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Belgrade and University of Novi Sad, in 1993 and 1997, respectively. He has created distinguished curricula in acoustics, audio signal processing, and speech technologies at FTS-UNS, which have attracted interest of several universities from the region who invited prof. Delić as a visiting professor. He has been leading major national and international research proje cts in the field of speech technologies in Serbia, and is the leader of the most renowned scientific research team in the field of speech technology in Western Balkans. Prof. Delić has large experience and skills in signal processing research and transfer to ICT applications, he has published nearly 300 scientific papers, and his research has evolved into a number of technical solutions widely applied across the region. For his contribution to innovation in the field of speech technology, Prof. Delić has received several prestigious awards. He is a member of international organizations such as IEEE and Audio Engineering Society (AES).

Audio-Visual Automatic Speech Recognition Using PZM, MFCC and Statistical Analysis

Saswati Debnath, Pinki Roy *

National Institute of Technology, Silchar, Assam (India)

Received 3 March 2020 | Accepted 15 October 2020 | Published 1 September 2021

LA UNIVERSIDAD EN INTERNET

Abstract

Audio-Visual Automatic Speech Recognition (AV-ASR) has become the most promising research area when the audio signal gets corrupted by noise. The main objective of this paper is to select the important and discriminative audio and visual speech features to recognize audio-visual speech. This paper proposes Pseudo Zernike Moment (PZM) and feature selection method for audio-visual speech recognition. Visual information is captured from the lip contour and computes the moments for lip reading. We have extracted 19th order of Mel Frequency Cepstral Coefficients (MFCC) as speech features from audio. Since all the 19 speech features are not equally important, therefore, feature selection algorithms are used to select the most efficient features. The various statistical algorithm such as Analysis of Variance (ANOVA), Kruskal-wallis, and Friedman test are employed to analyze the significance of features along with Incremental Feature Selection (IFS) technique. Statistical analysis is used to analyze the statistical significance of the speech features and after that IFS is used to select the speech feature subset. Furthermore, multiclass Support Vector Machine (SVM), Artificial Neural Network (ANN) and Naive Bayes (NB) machine learning techniques are used to recognize the speech for both the audio and visual modalities. Based on the recognition rate combined decision is taken from the two individual recognition systems. This paper compares the result achieved by the proposed model and the existing model for both audio and visual speech recognition. Zernike Moment (ZM) is compared with PZM and shows that our proposed model using PZM extracts better discriminative features for visual speech recognition. This study also proves that audio feature selection using statistical analysis outperforms methods without any feature selection technique.

Keywords

Audio-visual Speech Recognition, Lip Tracking, Pseudo Zernike Moment, Mel Frequency Cepstral Coefficients (MFCC), Statistical Analysis, Incremental Feature Selection (IFS).

DOI: 10.9781/ijimai.2021.09.001

I. Introduction

Human speech production and perception are bi-modal, like audio, visual features can also be extracted from speech. Research in Audio-Visual Automatic Speech Recognition (AV-ASR) is promising when a speech signal is affected by acoustic noise, different environments, and recording channels [1]. Speech is one of the ancient ways to express ourselves and speech recognition develops the methodologies that enable recognition of spoken word into text. There are many real-world applications where speech recognition is applied to authenticate [2], [3], especially for remote access of a system [4]. Audio alone also gives a good performance in a clean environment, but in a noisy environment, the signal may degrade [1]. Therefore, adding visual features make the system more robust, because visual features are less sensitive to noise [5]. But visual speech recognition is a challenging problem because visual features provide very less information as compared to an acoustic signal [5]. Research is going on in this area to find more and more robust and specific features that convey more accurate visual features. Visual features can be appearance-based, shape-based, and appearance and shape features. All the visual feature extraction methods include determination of the region of interest (ROI), face detection, and lip tracking. The audio-visual integration mechanism also plays a very crucial part in the AV-ASR research where two types of fusion can be done, feature fusion and decision level fusion [6]. Decision fusion is useful for individual analysis of time frames and phone segments. Frame level or feature level fusion is difficult because of frame mismatch and asynchrony of audio-visual data [6]. The fusion of audio-visual modality ensures better and convenient recognition than a single modality. AV-ASR can be applied to build a robust and secure authentication system, silent speech recognition system for deaf people, etc. But it introduces the challenging task of localization of mouth and lip tracking. Prashant Borde et al [5] have introduced the application of shape-based visual features for isolated word recognition. They have used Zernike Moment (ZM) [7] for visual feature extraction. This paper is primarily focused on building a speech recognition model that utilizes both audio and visual features i.e. audio-visual speech recognition based on audio-visual features and integration method. Here, we have used the shape-based visual features and the features are extracted from the lip contours.

The major contributions of this paper are as follows:

- Proper articulation is the most important lip-reading condition i.e. quality of speech of a speaker and angle of view. Thus, Pseudo

[*] Corresponding author.

E-mail addresses: dnsaswati@gmail.com (S. Debnath), pinkiroy2405@gmail.com (P. Roy).

Zernike Moment (PZM) [8] is proposed here for the extraction of visual features from the lip contour. The proposed algorithm extracts the shape based visual features to calculate the lip geometry of a speaker.

- Mel Frequency Cepstral Coefficients (MFCCs) [9] are widely used cepstral features extracted from the audio signal. In this study, the significance of MFCCs is calculated using different statistical algorithms. ANOVA, Kruskal-wallis, and Friedman tests are used in the proposed model to analyze different cepstral features and their significance. After the statistical analysis of features, the IFS method is used to select the features subset from the speech signal incrementally.
- To meet the final objective of AV-ASR, this paper proposes threshold-based decision fusion which improves the system performance.

Comparison of results is also explained in this paper based on the research paper published by Prashant Borde et al. Visual speech features are extracted using PZM and ZM using vVISWa [12] and 'CUAVE' [49] datasets and the results are compared for both the feature extraction techniques. Similarly, this paper also compares the results of audio speech recognition. The paper is organized as follows: Section II gives the literature review of AV-ASR, the proposed model is introduced in section III. Database description and experimental results are given in section IV and V. In section VI, we conclude our paper.

II. Literature Review

A. Audio-visual Speech Recognition

Audio-visual speech recognition is an active research area. To improve recognition performance in noisy environments visual information is added to automatic speech recognition.

Prashant Borde et al. [5] in 2014 have introduced Zernike features for visual speech recognition. The work described audio-visual speech recognition, which included face as well as lip detection, visual feature extraction, audio feature extraction, and recognition. The system was divided into two phases- the recognition of visual speech and the recognition of audio speech. Viola-Jones algorithm has been used for mouth localization or ROI detection. After extraction of ROI, the authors have used Zernike Moments (ZM) and Principal Component Analysis (PCA) for visual speech recognition. For audio speech recognition, they have used MFCC features. However, ZMs are sensitive to noise, and extracted features are scale as well as rotation variant.

Kuniaki Noda et al. [13] proposed AV-ASR, using a deep learning architecture, and introduced a connectionist-HMM. The system has three phases, in the first and second phase, deep de-noising autoencoder, as well as Convolutional Neural Network (CNN), have been used for acquiring noise-robust audio feature and for visual feature extraction, respectively. After that, a multi-stream HMM has been utilized here for integrating individual audio-visual features. De-noised MFCC features were used as an audio feature while CNN was used to predict the phoneme levels from the corresponding mouth area of the input image. After feature extraction, both the audio and visual features have been provided as an input to the Multi-stream HMM (MSHMM) for integration, which leads to a recognition of isolated words. However, the visual speech features extracted by CNN are not translation, rotation, and scale-invariant. Thus, the proposed method failed to meet the robustness due to illumination variation.

An experiment on continuous Audio-visual recognition was performed by Jeffrey B.Mulligan et al. [14]. They deployed the N-best approach for decision fusion. The recognizer that has been developed so far gives the best result in noise-free environments, but results degrade when it comes under noisy conditions. The authors have shown that their proposed system improves the robustness in all the situations where the audio signal is distorted. Data from both the audio-visual modality was first processed separately and then they combined them.

Visual speech information from the speaker's mouth region has been successfully shown to improve noise robustness of automatic speech recognizer by Gerasimos Potamianos et al. [15]. Thus, it has been promising to extend the usability into the human-computer interface. The authors have designed the visual front-end, based on a cascade of linear image transform. They have also added audio-visual speech integration. New work on a feature and decision fusion combination, the modeling of audio-visual speech asynchrony, and incorporating modality reliability estimates to the bi-modal recognition process have been analyzed. They also briefly touched upon the issues of audio-visual speaker adaptation. The experiments were carried out using three multi-subject bi-modal databases, ranging from small to large vocabulary recognition tasks, recorded at both visually controlled and challenging environments.

Namrata Dave [16] in 2015 has presented a lip-localization based visual feature extraction method. The proposed method segments the lip region from the image. To synchronize the lip movements with input audio they have segmented the lip region. Thus, the author has presented a color-based approach for the localization of lips. The main goal of their work was to synchronize lips with the input speech. Therefore, synchronizing with audio, viseme visual features have been extracted from the input video frame. HSV and YCbCr color models along with various morphological operations have been used. However, color-changing features are not very effective in AV-ASR research because they are sensitive to noise and illumination. Poor illumination does not give very good performance in a color model. Illumination affects the pixels values of an image. The color model also increases the experimental complexity.

Alin G et al. [17] proposed lip geometry and optical flow for capturing mouth movement. The method combined appearance-based features with the statistical approach for lip reading. However, the audio-only speech recognition has still lacked in robustness issues in a noisy condition while the video information is more reliable in real-time. The optical flow analysis captured the motion information of the speaker's mouth region. For the classification, they have used the Hidden Markov Model (HMM). A different noisy environment is a strong requirement for developing a robust speech recognition system. In this proposed method, the appropriate weights measurement is a very crucial part for different data medium. The author also mentioned that the system's accuracy could decrease because of a large number of features.

The lip movement of an individual speaker has been added to the acoustic features of speech for AV-ASR. Stéphane Dupont and Juergen Luettin [18] proposed a system that consists of three modules: the visual module, an acoustic module, and a sensor fusion module. Lip contour and grey level information were used as visual speech features. The acoustic features Perceptual linear prediction (PLP) and noise-robust RASTA PLP have been extracted from the speech signal. The system combined the visual and audio features using a multistream HMM. The appearance-based model for noise-robust audio-visual speech recognition has been introduced by the authors.

Continuous audio-visual digit recognition using N-best decision fusion has been introduced by Georg F. Meyer et al. [6]. The main contribution of the paper was decision fusion in audiovisual continuous speech recognition at the utterance level and proposal of an algorithm called N-best decision fusion. For the audio feature, they have taken 12 orders cepstral coefficients (MFCC) and calculated the word error recognition (WER) rate. For video feature recognition, lip shape has been measured.

In [45], the authors introduced visual speech recognition by calculating the Gray Level Co-occurrence Matrix (GLCM) and Gabor convolve algorithm for discriminative feature extraction of the lip. They have collected a dataset of three Indian languages of English, Kannada, and Telugu for the experiments. GLCM provides the statistical texture features of lip movements. In this work, the authors have used four GLCM features such as contrast, energy, entropy, and correlation for the calculation of lip parameters. The mean and variant of the filtered image have also been calculated by using the Gabor filter. Thus, the main objective of this work was to analyze the texture of different images of lip movements.

B. Audio Speech Recognition

The work in [19] was carried out for the automatic recognition of English digits in 2010. The main objective of this study was to design and execute an English digits recognition system with the help of Matlab; using Hidden Markov Model. The framework perceives the speech waveform by making an interpretation of the speech waveform into an arrangement of high- light vectors using the popular technique MFCC.

Hindi Number Recognition [20] system was carried using Gaussian Mixture Models and MFCC. In the primary stage vowel acknowledgment models are created, which is supervised learning and in the subsequent stage, testing of the prepared models has been performed. Spectral components are separated from the discourse signals of the digits (0-9) and these elements are utilized to prepare Gaussian mixture models.

In [21] an idea is proposed that was a digit recognition system using Reservoir Computing (RC) which is a concept of machine learning. It is a non-linear dynamical system. It computes the state likelihood in HMM through two-layer Recursive Neural Network. The input hidden layer repetitively interfaces non-straight neurons with a settled number of non-prepared coefficients which is called a store (reservoir). They tested multilayer systems with 8000 and 16000 neurons. Later they performed a systematic evaluation using AEF (Advanced front end) where they replaced MVN features with AEF features and obtained significant gains in GMM-HMM recognizer.

S. Lokesh et al. [22] discovered a bidirectional recurrent neural network-based automatic Tamil speech recognition system in 2018. Bidirectional recurrent neural network (BRNN) with a self-organizing map (SOM) is used for the classification of Tamil speech. Savitzky–Golay filter is used for pre-processing to remove noise. For feature extraction, they have used discrete cosine transform and perceptual linear predictive coefficients. Using their proposed BRNN-SOM method 93.6 % accuracy was achieved for Tamil speech recognition.

The selection of feature vector from MFCC and Sequence-based Mapped Real Transform (SMRT) coefficients has been proposed by the author Mini p p et al. [23]. The first feature set was the coefficients extracted from all frames and after feature fusion, feature dimension reduction has carried out using a statistical measure such as energy, sum, mean, standard deviation, and energy distribution. These statistical measures are applied on the time average base to derive the second feature set. To solve the length variation problem of the speech signal, all the statistical measures are applied on the ensemble average base for generating the third feature vector. Furthermore, they have used Support Vector Machine (SVM) for the classification of the speech signal.

In [24] the authors introduced optimal speech feature extraction as well as feature selection using Artificial Bee Colony and Particle Swarm Optimization (ABC-PSO) hybrid algorithm. They have extracted eight types of statistical and acoustic features and ABC-PSO has been proposed for the selection of optimal features. After that SVM has been used to carry out the recognition process.

Nasir Saleem et al. [46] presented a detailed survey on unsupervised single-channel speech enhancement algorithms. The speech enhancement algorithms on unsupervised single-channel perspectives are analyzed and presented. Various methods have been discussed by the author for improving noisy speech. They have reviewed different approaches such as spectral subtraction, wiener filtering, minimum mean square error estimators, signal subspace, etc, and presented the experimental overview of these approaches. The authors have found that these methods show improvement in speech quality but speech intelligibility remains medium, thus, various problems have been introduced in the paper for designing robust single-channel speech enhancement algorithms.

In 2019, Nasir Saleem et al. [47] proposed speech enhancement using deep neural network (DNN) in complex noisy environments. They have also used an ideal binary mask (IBM) as a binary classification function during training and the trained DNNs are used for estimating IBM during the enhancement stage. The mean square error (MSE) has been used as an objective cost function at various epochs. The experimental results at different input SNR of this research showed that DNN-based speech enhancement performed better in a complex noisy environments than the competing methods in terms of perceptual evaluation of speech quality (PESQ), segmental signal-to-noise ratio (SNRSeg), log-likelihood ratio (LLR), weighted spectral slope (WSS), short-time objective intelligibility (STOI) and also improved the speech intelligibility an average 6.5% improvement during experiments.

Issues from the literature review:

- In the visual speech recognition approach, features are not translation, rotation invariant in many studies.
- Audio-visual integration is not done in many research works.
- Shape-based feature ZM has minimum feature dimensions also is very sensitive in noise.
- In audio, there are very few algorithms used to select the speech features, the majority of work focused on feature extraction and classification using machine learning.
- In the fusion method, a frame-level fusion mismatches the audio and visual frame.
- Feature fusion is partially valid because there is a different data rate of audio and visual data and differing segmentation [6].
- Many lip-reading systems use vizeme based representations for the visual recognition and phone-based for audio recognition but the co-articulation effects cause the asynchrony between the phone level segmentation of audio and visual data.

III. Proposed Methodology

Audio-visual speech recognition includes two separate processes of recognition: audio speech recognition and visual speech recognition. The proposed methodology of audio-visual speech recognition is shown in Fig. 1. The system includes the following steps.

a) **ROI detection**: Face and ROI have been detected using Viola Jones algorithm.

b) **Visual speech feature extraction**: After ROI detection visual features are extracted using PZM from the lip contour. PZMs are rotation and translation invariant of the image. ZM is also used with our dataset and shows that our proposed model using PZM gives better recognition than the ZM.

c) **Audio feature extraction**: MFCC feature extraction is used and also extraction of significant features is done using statistical analysis.

d) **Audio feature selection**: Statistical algorithms are used to select efficient audio features. The proposed method is a combination of different statistical analysis and IFS.

e) **Classification**: Classification of audio-visual speech is carried out using multiclass SVM, ANN and NB classifier.

f) **Decision making**: Combined decision is taken from audio and visual speech recognition. Here, we propose a threshold-based decision level fusion to overcome frame level mismatch.

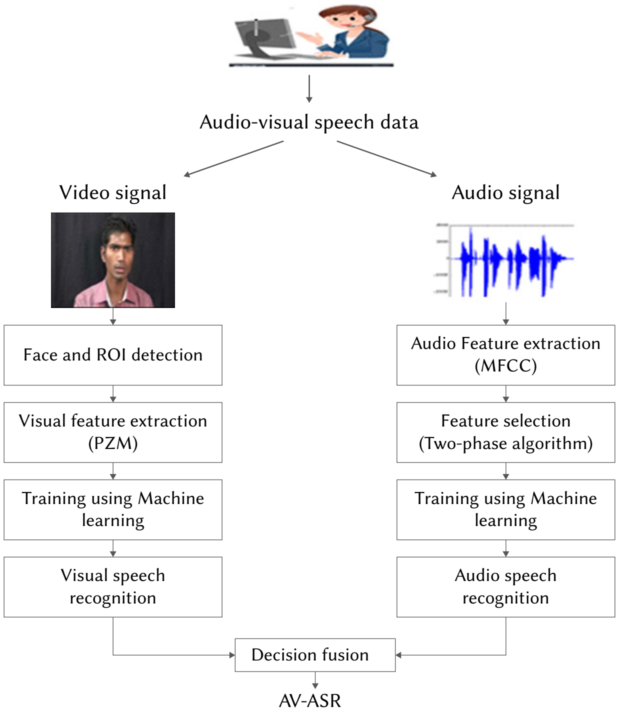

Audio-visual speech data
Video signal
Audio signal
Face and ROI detection
Visual feature extraction (PZM)
Training using Machine learning
Visual speech recognition
Audio Feature extraction (MFCC)
Feature selection (Two-phase algorithm)
Training using Machine learning
Audio speech recognition
Decision fusion
AV-ASR

Fig. 1. Proposed model of AV-ASR.

A. Visual Feature Extraction

Visual features can be categorized by appearance-based, shape-based, and appearance and shape-based features. All of these methods include the determination of the ROI, face detection, and lip tracking. For visual feature extraction, we consider the video stream as an input. From each utterance, we have extracted the frames and processed each frame separately to obtain the discriminative features. Visual feature extraction method includes:

- Detection of the face and ROI (speaker's lip contour) using Viola-Jones algorithm.
- Calculate the visual features from lip contour.

1. Viola-Jones For ROI Detection

The Viola-Jones object detection framework facilitates Haar-Like [25] [26] features to be extracted from a face image as the initial step. The reason for using Haar-like features over the raw pixel value of the image is to reduce the in-class variability while increasing the out-of-class variability, which makes the classification easier. The contrast variances between the pixel groups are used to determine relative light and dark areas. It considers neighbouring rectangular regions in the image that is targeted for facial detection. After that, it sums up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image. In a human face, it is common that the region of the eyes is darker than the region of the cheeks. Therefore, a common Haar-like feature for face detection is a set of two adjacent rectangles that lie above the eye and the cheek region [27]. Fig. 2 and Fig. 3 depict pre-processing of visual feature extraction method.

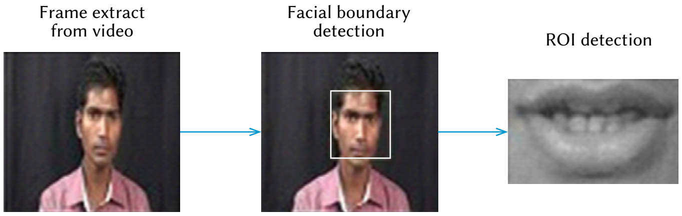


Fig. 2. Pre-processing of visual feature extraction from video.

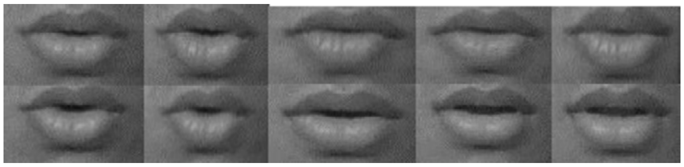


Fig. 3. Lip contour: open and close mouth for a particular word uttered by speaker.

B. Visual Speech Feature Extraction

Zernike Moment (ZM)

ZM is an orthogonal polynomial which is independent of scale and rotation of the image. It has less information redundancy and is used to capture discriminating feature of image frames.

Pseudo Zernike Moment (PZM)

PZMs are orthogonal moments on the unit disk defined by mapping an image onto a set of pseudo-Zernike polynomials [8]. PZMs are also rotation and flipping invariant. ZM polynomials are defined in polar coordinates. The orthogonal moments represent an image with the minimum number of redundant information [8]. PZM of order n and repetition of m can be computed over a unit disk by the following equation [28], [29].

$$A_{n,m} = \frac{n+1}{\pi} \iint_{x^2+y^2\leq 1} f(x,y) V^{*}{}_{n,m}\,(x,y)dxdy \tag{1}$$

Where, n= 0, 1, 2, 3, …., ∞ defines the order, f(x,y) is the function being described, * denotes the complex conjugate, while m is the positive or negative integer depicting the angular dependence, $V^*_{n,m}(x,y)$ is the complex pseudo Zernike polynomial which is defined by:

$$V_{n,m}(x,y) = R_{n,m}(x,y)\, \exp^{jm\theta} \tag{2}$$

Where (x,y) are defined over the unit disc, here, $R_{n,m}(x,y)$ is the real-valued radial polynomial, n ≥ 0, |m|≤ n, and the radial polynomials $R_{n,m}(x,y)$ are defined as:

$$R_{n,m}(x,y) = \sum_{s=0}^{n-|m|} (-1)^s \frac{(2n+1-s)!\,(x^2+y^2)\frac{n-s}{2}}{s!\ [n+|m|+1-s]!\ \ [n-|m|-s]!} \tag{3}$$

Where, $n = 0, 1, 2, 3,$ …., ∞ defines the order and m is the positive or negative integer value subject to |m|≤n. According to simple enumeration, the set of pseudo-Zernike polynomials contains $(n + 1)^2$ linearly independent polynomials of degree ≤n [28].

PZM has been proven to be more efficient than the conventional ZM because of their feature representation capabilities [29]. The feature vector of ZM has 36 dimensions for maximum 10th order while PZM has 66 dimensions of feature vectors [29]. PZM is more efficient to recognize the similar image frames since the number of features are more in PZM than the ZM. Also, it has been proven by Mukundan et al. that PZM are less sensitive to noise than the ZM for recognition of the image frame [30], [31]. The proposed visual feature extraction method using Viola-Jones and PZM is given in algorithm 1.

Algorithm 1: Shape-based visual features calculation using PZM

```
Input: Video of a speaker
Output: Lip geometry calculation
procedure: VISUAL SPEECH FEATURE EXTRACTION
    Extract frame // Read video data
    N←number of frames
    bbox←(facedetector,N)
    videoFrameN = insertShape(videoFrame,j Rectanglej, bbox);
    image write "detectedface.jpg"
    MouthDetect : I←vision.
    CascadeObjectDetector('Mouth','MergeThreshold'k);
    k=threshold value [detection of object using Viola -Jones]
    I← detected mouth area from each frame
    for i= 1 to I do
        img ←Load image
        Normalize co-ordinates to [-squreroot2,squreroot2] and
        calculate origin or centroid ; x and y two coordinates
        x1 = 2/( squreroot (2)*(d(1)-1)); d= size of image (dimension)
        y1 = 2/( squreroot (2)*(d(1)-1));
        [x,y]=meshgrid(1/ squreroot (2):x1:1/ squreroot (2),1/
        squreroot (2: -y1:
        -1/ squreroot (2));
        $x^2 + y^2 \le 1$ ; // Compute unit circle
        pixels inside the unit circle [cimage, cindex] = p1(img, m);
        m=zeros of d
        z = p1(x+i*y,m);
        p ← compute z;
        q ← angle(z);
        Compute order n and repitition m
        for n= 1:length (l) do //n=order of PZM
            n1 = l(n);
            for r=1:length(m1) do
                V = pzpolynomial(n1, m1(r), p, q);
                PZp1 = cimage * conj(V );
                $PZM = 2 * (n1 + 1) * sum(sum(pzp1))/(d(1)^2 * pi)$;
                pzm (u,z)= round(p(PZMoment(I,i,j)));
                Magnitude of each component is evaluated and
                rounded off to nearest integer.
            end for
        end for
    end for
end
```

C. Audio Feature Extraction and Statistical Analysis

In this step, 19 MFCCs are extracted from each input audio data. Here, we have used 19 MFCC features because the increasing level of spectral information comes from the higher order of coefficients and we need more information from more coefficients to select efficient features. The main aim of feature selection is to select the effective feature subset that can increase the classification accuracy while reducing the irrelevant and redundant features [32]. The feature extraction and selection methods are described below:

MFCC:

The most popular acoustic feature extraction technique MFCC was first introduced in 1980 by David and Mermelstein [9]. Today most of the speech recognition system focuses on the short-term spectral features which are captured from a short frame of the speech signal. The MFCC feature extraction consists of the following major steps:

- Framing and windowing: At first, each speech signal breaks down into short time duration by splitting the signal into several frames instead of analyzing the complete signal at once. After framing the signal, a window function is multiplied with each frame of the speech signal. We perform windowing in order to avoid unnatural discontinuities in the speech segment and the distortion in the underlying spectrum.
- Discrete Fourier Transforming: For extracting spectral information from a discrete frequency band Discrete Fourier Transform (DFT) is used. Fast Fourier Transform (FFT) is the most commonly used algorithm to compute the DFT. Here we are using FFT to convert the signal from time domain to frequency domain for preparing the next stage (Mel frequency warping).
- Mel-Frequency Warping: The FFT of the signal gives the magnitude, frequency response of each frame. After getting FFT the Mel filter bank includes the following calculations.
- The Mel scale: The result of FFT is the information about the amount of energy that the signal contains at each frequency band. For a given frequency f we can use the following formula to compute the Mels in Hz: Mel scale [9] is defined as:

  $$m_f = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{4}$$

  Low frequency components of the speech signal carries much more information compared to the high frequency components. Mel scaling is performed in order to place more emphasis on the low frequency components. Since Mel filter banks are non-uniformly spaced on the frequency axis, so we have more filters in the low frequency regions and less number of filters in high frequency regions.
- Cepstrum: In the final step, the log mel spectrum is converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs) [33]. For the given frame analysis, the cepstral representation of the speech spectrum is a good representation of the local spectral properties of the signal. Since the mel spectrum coefficients are real numbers (and so are their logarithms), Discrete Cosine Transform (DCT) is performed to convert them into time domain [34].
- As per the procedure above, for each speech frame of about 20 ms with overlap of about 10 ms, a set of mel-frequency cepstrum coefficients are computed. This set of coefficients is called an acoustic vector. These acoustic vectors are used to represent and recognize the speech. Therefore, each input utterance is transformed into a sequence of acoustic vectors. We have extracted 19 mfc coefficients for each frame. From these 19 coefficients, feature 1 is the energy value of speech and feature 2 represents the broad shape of spectrum, features 3 to feature 7 represent the pitch information, features 8 to 13 represent the shape of spectrum and 14 to feature 19 provide the pitch or tone information which are the lowest dimensions of DCT coefficients. All these features are not equally important; therefore, we have calculated the F-statistics using statistical algorithm. The highest F-value of feature represents the most significant speech feature.

D. Audio Feature Selection

Analysis of Variance (ANOVA): ANOVA is a very effective statistical method to test the difference in means between groups [10]. It assesses the potential difference in a scale-level dependent variable to a nominal-level variable having two or more categories. There are many advantages for that ANOVA has been used for feature selection [35], [36].

$$F(\xi) = \frac{sB^2(\xi)}{sW^2(\xi)} \tag{5}$$

Where, $sB^2(\xi)$ is the sample variance between groups (also called Mean Square Between, MSB) and $sW^2(\xi)$ is sample variance within groups (also called Mean Square Within, MSW) [16]. Decision is made using F-statistics [16], in one way ANOVA, the F-statistics is the ratio calculated by following equation:

$$F = \frac{variation\ between\ sample\ means}{variation\ within\ the\ samples} \tag{6}$$

Audio feature selection (ANOVA with IFS):

The method is calculated using the following steps:

ANOVA calculation:

- First, find the group mean of MFCC feature of each input data group.
- Find the overall mean i.e. mean of all MFCC feature matrices.
- Calculate within group variation, that is, the total deviation of each feature score from the group mean of each input data.
- Calculate the deviation of each input group mean from the overall mean ($ m2 $), i.e. the between group variation.
- Find the F statistics, the ratio between group variations to within group variation. The ratio calculates the measure of dispersion i.e. how much difference is there from each feature to the mean. Larger the value of F defines the more significant speech feature. These steps are performed for all the speech features.

IFS calculation for cepstral features is as follows:

- The feature subset starts from the feature with the highest F value in the ranked feature set.
- A new feature subset is produced when the feature with the second-highest F value is added.
- Until all the candidate features are added, this process continues from the highest F value to the lowest F value.

Kruskal-wallis test: Kruskal-wallis is a rank based nonparametric test which is used to evaluate the significant difference between two or more groups of an independent variable on an ordinal or continuous dependent variable. It is a Chi-square distribution. The test is performed to obtain the Chi-square value of each feature and rank those feature in a decreasing order. Null hypothesis is important for this test, if the Chi-square distribution score is less than the critical chi-square value then null hypothesis is accepted or has same group otherwise it is rejected or in a different group [37].

- Chi-square distribution uses categorical data, the mean (μ) of the distribution is the number of degrees of freedom (f) i.e. μ= f and variance (σ) of the distribution is σ=2 * f.
- The Chi-square is calculated using the following equation:

$$Chi - square = \sum \frac{(O - E)^2}{E} \tag{7}$$

Where, O is the observed value, E is the expected value.

- It calculates whether there is any statistically significant difference between different coefficients of cepstral features of different classes. The less Chi-square value indicates that less statistical difference. The higher Chi-square value indicates the more statistical difference between cepstral features of different classes.

Therefore, the features with higher Chi-square values are selected using this Chi-square calculation. After feature ranking IFS is used here to select features for the classifier. The feature with highest Chi-square value is selected first, and then second highest value and so on.

Friedman test: Friedman test is a non-parametric test and alternative to one-way ANOVA with repeated measures. It is used to test differences between groups when the dependent variable being measured is ordinal [38]. The probability distribution is a Chi-square distribution and according to Chi-square value, features are ranked in order. Although Kruskal-wallis and Friedman test both calculate the Chi-square value but difference is there. Kruskal-wallis is the non-parametric test equivalent to one-way ANOVA while Friedman is a non-parametric test equivalent to two-way ANOVA.

E. Classifiers

Multiclass classifier SVM:

SVM [11], [39], [42] is a supervised machine learning that estimates decision surfaces directly rather than modeling a probability distribution across the training data. The kernel functions of SVM can be linear, Radial Basis Function (RBF), and polynomial function, which are used in this experiment. In a characteristics way, SVMs are two-class classifier but it can be used for multi-class classification problems. The methods used for multi-class classification are one-versus-all and one-versus-one. In this experiment, we have used the one-versus-all method for multi-class classification. We have 10 classes in this experiment, thus we generate 10 binary classifiers for 10 respective classes. After that, class levels are generated and the training dataset is created for each classifier of each class. After training, we have passed the test dataset. If the input test data belongs to a particular class, then the classifier generates a positive response and all the other classifiers provide a negative response.

Artificial neural networks (ANN): ANN [39], [40] is used for pattern classification because of its capability of nonlinear, non-parametric relationships between input data and output. Multi-layer feed-forward neural networks are the most popular neural networks which are trained using backpropagation learning algorithm. In this work, a multi-layer feed-forward ANN with the sigmoidal activation function is used with different hidden units.

Naive Bayes (NB): In NB [41], 'kernel' and 'normal' functions are used to model the feature distribution to decide the best distribution, and performed testing using fourfold cross-validation. It calculates the probability using Bayes theorem [41].

The audio-visual features are fed into ANN, SVM and NB classifiers to classify the speech. The performance is calculated by total number of words correctly recognized during the testing phase.

$$RR = \frac{C_s}{T_s} \times 100\% \tag{8}$$

Where, RR denotes recognition rate and C_s and T_s represent the correctly identified test sample and total supplied test sample, respectively. Audio speech recogn ition accuracy is computed using the same method.

F. Decision Fusion

It is important to integrate audio and visual recognition to obtain the final objective of AV-ASR. For the integration of two types of fusion such as feature fusion and decision, fusion can be done. Decision fusion is useful for isolated word recognition, where each recognized token is considered as a decision. Here, we propose a threshold-based fusion of audio-visual speech at the decision level. Based on the recognition accuracy of the system we consider the threshold for each system. According to that threshold of the individual system, we calculate the system performance. If accuracy is greater than or equal to the threshold, then consider that input data is acceptable. When one of the recognition systems recognizes the respective input data of audio and visual speech, based on the threshold the system considers that speech gets recognized. In this work, we have processed audio and visual data separately. After the recognition of audio-visual speech, we have used decision fusion for the integration of audio-visual speech. Audio-visual speech has a different data rate and it creates a synchronization problem when encounters frame as well as feature level fusion [6]. Thus, to avoid the asynchrony of audio and visual data we have developed decision fusion. Many types of research have introduced a dynamic weighting method for AV-ASR integration. However, the decision fusion method improves performance while overcoming the issues of frame and feature level fusion. The decision fusion proposed in this paper is given in algorithm 2.

Algorithm 2: Decision fusion algorithm of audio-visual speech recognition

```
1. Input: Output of audio speech recognition and visual speech
          recognition
2. Output: Combined decision
3. Set threshold for audio speech recognition, and visual speech
   recognition
4 X = Audio speech recognition threshold
  Y = Visual speech recognition threshold
5 if (audio speech recognition ≥ X) || (visual speech recognition ≥Y )
  then
    recognition=1
    (accept) otherwise
    0 (reject)
end
```

IV. Dataset Description

We have used an audio-visual English digit database ‘vVISWa’ and ‘CUAVE’ for AV-ASR.

‘vVISWa’ dataset: Prashant Borde et al. [12] published a paper about ‘vVISWa’ dataset in 2016. English 10 digits (zero, one, two, three, four, five, six, seven, eight, and nine) have been recorded. The database has been recorded in a lab environment. Dataset consists of 10 speakers, 6 male, and 4 female. Each speaker uttered each word 10 times. So, one word is uttered 100 times by a different speaker. Each individual word is uttered without any head movement.

‘CUAVE’ dataset: E.K. Patterson et al. [49] published the details of ‘CUAVE’ audio-visual database for multimodal human-computer interface. The ‘CUAVE’ dataset is used here for the experiments and to compare the results. The dataset consists of ten English digits from 0 to 9 of 36 speakers (18 male and 18 female speaker). Each digit is uttered five times by each speaker; thus, the total 1800 words has been used. The database was recorded in an isolated sound booth at a resolution of 720 x 480 with the NTSC standard of 29.97 fps. The audio is 16-bit, stereo, at a sampling rate of 44 kHz. There is also word-level labelling at millisecond accuracy, done manually for all sequences of the database.

V. Experimental Result and Analysis

A. Visual Speech Recognition

After detecting the ROI, PZMs are calculated for the lip region. PZM measures how lips are moving for a particular speech and based on these visual features, we classify the speech spoken by the speaker. The steps of calculating lip movements using PZM are given below:

1. First, consider the open area of the mouth for each frame of the lip.
2. Take the origin of the lip contour.
3. The pixel coordinates of lips are normalized to the range of a unit circle. i.e. x2 + y2 ≤ 1
4. Calculate the angle and coordinates of each point. These are the features of visual speech.
5. Repeat steps 1 to 4 for every frame of a particular word and generates a feature matrix.

Calculate the angle and coordinates of each point. These are the features of visual speech. We have extracted frames from the video for each utterance. From each frame face and ROI is detected using the Viola-Jones algorithm. ZM and PZM [28] are calculated for every frame of lip contour to extract the discriminating feature of visual information. The PZMs have effective mathematical calculation to capture the different movements of the image. Here, we have taken 10 discriminative frames of lip contour and calculated the pseudo-Zernike feature of each frame. We have extracted 19 coefficients for each frame and 10 frames for a single utterance, therefore 10x19 feature matrix for a single digit. After extracting the visual features, we have applied multiclass SVM, ANN, and NB machine learning to train the system. The performance is calculated by the total number of words correctly recognized using visual features during the testing phase. The performances of visual speech recognition with different classifiers are presented in Table I, Table II, and Table III using ‘vVISWa’ dataset.

We have carried out all the experiments using ‘CUAVE’ dataset also and presented the recognition rate in Table IV, Table V and Table VI for visual speech recognition using different classifier.

TABLE I. Visual Speech Recognition Using ZMs, PZMs and ANN for ‘vVISWa’ Dataset

Exp. No.	Hidden layer	No. hidden nodes	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	2	30,20	70.00	73.34
2	2	40,30	69.54	72.98
3	2	50,40	69.23	71.56
4	2	60,50	70.12	72.89
5	2	70,60	70.00	72.10

TABLE II. Visual Speech Recognition Using ZMs, PZMs and NB for ‘vVISWa’ Dataset

Exp. No.	Distribution function	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	Normal	68.20	72.00
2	Kernel	70.34	74.65

TABLE III. Visual Speech Recognition Using ZMs, PZMs and SVM for ‘vVISWa’ Dataset

Exp. No.	Kernel function	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	**Radial basis function (RBF)**	**68.00**	**75.23**
2	Linear	61.54	67.45
3	Polynomial	66.12	70.56

TABLE IV. Visual Speech Recognition Using ZMs, PZMs and ANN for 'CUAVE' Dataset

Exp. No.	Hidden layer	No. hidden nodes	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	2	30,20	72.15	73.44
2	2	40,30	68.24	73.08
3	**2**	**50,40**	**73.12**	**75.34**
4	2	60,50	72.27	73.90
5	2	70,60	71.00	73.00

TABLE V. Visual Speech Recognition Using ZMs, PZMs and NB for 'CUAVE' Dataset

Exp. No.	Distribution function	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	Normal	67.20	70.00
2	**Kernel**	**73.34**	**75.00**

TABLE VI. Visual Speech Recognition Using ZMs, PZMs and SVM for 'CUAVE' Dataset

Exp. No.	Kernel function	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	**Radial basis function (RBF)**	**72.15**	**76.03**
2	Linear	64.00	67.54
3	Polynomial	65.21	71.60

B. Audio Speech Recognition

We have extracted19-dimensional MFCC features for the experiment. The statistical test is carried out to rank the MFCC features based on the F-statistics which has been discussed in section III D. The feature with the highest F value is placed first, that is ranked one, followed by the feature with second-highest value that is ranked two, and so on. The F-statistics is calculated by equation (6). The performance of the system is measured using equation (7). We have calculated the recognition rate using all the MFCC features and also the features subset resulting from the feature selection method. The recognition rate is carried out using SVM, ANN, and NB and IFS gradually concatenates the features for all the classifiers.

Table VII, Table VIII, and Table IX show the F-statistics value of speech features after statistical analysis for 'vVISWa' dataset. The performance of the system after feature selection technique using SVM, ANN, and NB are depicted in Table X, Table XI, and Table XII. Table XIII shows the performance of audio speech recognition using MFCC. From the experiment, it has been observed that using 'vVISWa' dataset, the SVM classifier with kernel function 'RBF' gives the highest accuracy that is 96.42 % for 12 cepstral features. The highest accuracy is obtained using ANOVA with IFS feature selection method. All the experiments are carried out using 'vVISWa' dataset.

TABLE VII. F-value After Statistical Analysis of MFCC Using ANOVA

Feature set	F-value	Feature set	F-value
f1	719.63	f11	127.68
f2	477.40	f12	31.50
f3	249.84	f13	65.62
f4	253.10	f14	110.67
f5	249.15	f15	35.55
f6	115.42	f16	85.87
f7	154.19	f17	50.45
f8	22.15	f18	21.76
f9	47.34	f19	45.56
f10	226.65		

TABLE VIII. F-value After Statistical Analysis of MFCC Using Kruskal-wallis

Feature set	F-value	Feature set	F-value
f1	634.64	f11	127.68
f2	376.00	f12	22.10
f3	265.23	f13	47.65
f4	364.78	f14	58.11
f5	188.25	f15	25.62
f6	60.12	f16	19.27
f7	91.19	f17	42.55
f8	62.15	f18	20.27
f9	32.84	f19	18.45
f10	87.65		

TABLE IX. F-value After Statistical Analysis of MFCC Using Friedman Test

Feature set	F-value	Feature set	F-value
f1	671.34	f11	127.68
f2	423.37	f12	46.20
f3	286.23	f13	35.75
f4	364.78	f14	22.87
f5	210.15	f15	29.56
f6	92.32	f16	17.68
f7	64.99	f17	28.34
f8	63.65	f18	16.07
f9	56.34	f19	16.35
f10	58.75		

TABLE X. Audio Speech Recognition Rate Using MFCC, Feature Selection and SVM for 'VVISWa' Dataset

Exp. No.	Feature selection method	No. of features	Kernel function	Accuracy
1	ANOVA+IFS	12	Radial basis function (RBF)	96.42
2	ANOVA+IFS	13	Linear	78.11
3	ANOVA+IFS	14	Polynomial	80.66
4	Kruskal-wallis+IFS	16	Radial basis function (RBF)	95.31
5	Kruskal-wallis+IFS	14	Linear	77.78
6	Kruskal-wallis+IFS	17	Polynomial	85.65
7	Friedman+IFS	12	Radial basis function (RBF)	93.45
8	Friedman+IFS	13	Linear	77.57
9	Friedman+IFS	14	Polynomial	90.34

TABLE XI. Audio Speech Recognition Using MFCC, Feature Selection and ANN for 'VVISWa' Dataset

Exp. No.	Feature selection method	No. of features	Hidden layer and nodes	Accuracy
1	ANOVA+IFS	13	2 (30,20)	92.06
2	ANOVA+IFS	11	2 (40,30)	94.78
3	ANOVA+IFS	14	2 (50,40)	90.88
4	ANOVA+IFS	13	2 (60,50)	93.96
5	ANOVA+IFS	13	2 (70,60)	90.34
6	Kruskal-wallis+IFS	12	2 (30,20)	92.18
7	Kruskal-wallis+IFS	13	2 (40,30)	91.68
8	Kruskal-wallis+IFS	13	2 (50,40)	91.85
9	Kruskal-wallis+IFS	19	2 (60,50)	90.75
11	Friedman+IFS	10	2 (30,20)	89.67
12	Friedman+IFS	12	2 (40,30)	92.89
13	Friedman+IFS	16	2 (50,40)	93.17
14	Friedman+IFS	14	2 (60,50)	92.78
15	Friedman+IFS	19	2 (70,60)	91.56

TABLE XII. Audio Speech Recognition Using MFCC, Feature Selection and NB for 'VVISWa' Dataset

Exp. No.	Feature selection method	No. of features	Distribution function	Accuracy
1	ANOVA+IFS	15	Normal	93.72
2	ANOVA+IFS	13	Kernel	94.23
3	Kruskal-wallis+IFS	17	Normal	94.01
4	Kruskal-wallis+IFS	16	Kernel	94.57
5	Friedman+IFS	19	Normal	91.72
6	Friedman+IFS	15	Kernel	92.87

TABLE XIII. Audio Speech Recognition Using MFCC and SVM

Exp. No	Classifier	Accuracy (%) using 'vVISWa' dataset	Accuracy (%) using 'CUAVE' dataset
1	SVM ('RBF' kernel function)	93.86	94.55
2	ANN (Hidden layer-2 and hidden nodes- 50,40)	93.67	93.35
3	NB (kernel distribution function)	92.86	94.00

The 'CUAVE' digit dataset is also used for the audio speech recognition and recognition accuracies are presented in Table XIV, Table XV and Table XVI. The highest accuracy achieved by this dataset is 98.0 % for 13 number of features using ANOVA with IFS.

TABLE XIV. Audio Speech Recognition Using MFCC, Feature Selection and SVM for 'CUAVE' Dataset

Exp. No.	Feature selection method	No. of features	Kernel function	Accuracy
1	**ANOVA+IFS**	**13**	**Radial basis function (RBF)**	**98.00**
2	ANOVA+IFS	12	Linear	75.23
3	ANOVA+IFS	13	Polynomial	82.12
4	Kruskal-wallis+IFS	15	Radial basis function (RBF)	96.32
5	Kruskal-wallis+IFS	14	Linear	77.00
6	Kruskal-wallis+IFS	16	Polynomial	82.02
7	Friedman+IFS	12	Radial basis function (RBF)	95.45
8	Friedman+IFS	13	Linear	76.57
9	Friedman+IFS	15	Polynomial	92.81

TABLE XV. Audio Speech Recognition Using MFCC, Feature Selection and ANN for 'CUAVE' Dataset

Exp. No.	Feature selection method	No. of features	Hidden layer and nodes	Accuracy
1	ANOVA+IFS	14	2 (30,20)	91.32
2	ANOVA+IFS	12	2 (40,30)	93.45
3	ANOVA+IFS	13	2 (50,40)	94.75
4	**ANOVA+IFS**	**13**	**2 (60,50)**	**96.55**
5	ANOVA+IFS	14	2 (70,60)	96.00
6	Kruskal-wallis+IFS	13	2 (30,20)	91.08
7	Kruskal-wallis+IFS	13	2 (40,30)	92.52
8	Kruskal-wallis+IFS	14	2 (50,40)	93.00
9	Kruskal-wallis+IFS	18	2 (60,50)	94.00
11	Friedman+IFS	10	2 (30,20)	90.22
12	Friedman+IFS	12	2 (40,30)	92.00
13	Friedman+IFS	15	2 (50,40)	93.55
14	Friedman+IFS	13	2 (60,50)	93.00
15	Friedman+IFS	19	2 (70,60)	92.11

TABLE XVI. Audio Speech Recognition Using MFCC, Feature Selection and NB for 'CUAVE' Dataset

Exp. No.	Feature selection method	No. of features	Distribution function	Accuracy
1	ANOVA+IFS	14	Normal	93.55
2	ANOVA+IFS	13	Kernel	95.74
3	Kruskal-wallis+IFS	16	Normal	93.00
4	Kruskal-wallis+IFS	15	Kernel	94.17
5	Friedman+IFS	17	Normal	90.00
6	Friedman+IFS	16	Kernel	93.65

C. Audio-Visual Speech Recognition Fusion

Here, we have considered decision level fusion for combining two systems because feature level fusion encounters the frame mismatch. The individual word recognition rate has been calculated for both audio and visual speech, after that using decision logic we integrate two modalities for the better result.

If one recognition system fails to recognize the input digit, then we can consider the result using another system. Decision fusion provides a better recognition rate for the overall system because each individual word is recognized as a token. The decision has been taken based on logic such that if the audio signal recognition rate is more than 90%, we have considered that audio speech is recognized. If the visual speech recognition rate is more than 70%, we have considered that visual speech is recognized. Thus, when the accuracy is greater than or equal to the threshold, then it is considered that the input data is acceptable. Based on the threshold, when one of the recognition systems recognizes the respective input data of audio and visual speech, the system considers that speech gets recognized. The proposed decision fusion method is represented by Algorithm 2.

D. Comparison of Results and Analysis

We have experimented separately for both audio and video data. The performances of our proposed model for audio speech recognition are 96.42 % and 98.00 % using 'vVISWa' and 'CUAVE' dataset respectively. For visual speech recognition, we have used lip tracking and 75.23 % accuracy is achieved using the proposed visual speech recognition model for 'vVISWa' dataset. For 'CUAVE' dataset, the recognition accuracy of visual speech is 76.03 %. We have compared the results of the existing model using ZM and our proposed model. Prashant Borde et al. [5] introduced ZM and MFCC features for audio-visual speech recognition respectively and achieved 63.88% accuracy

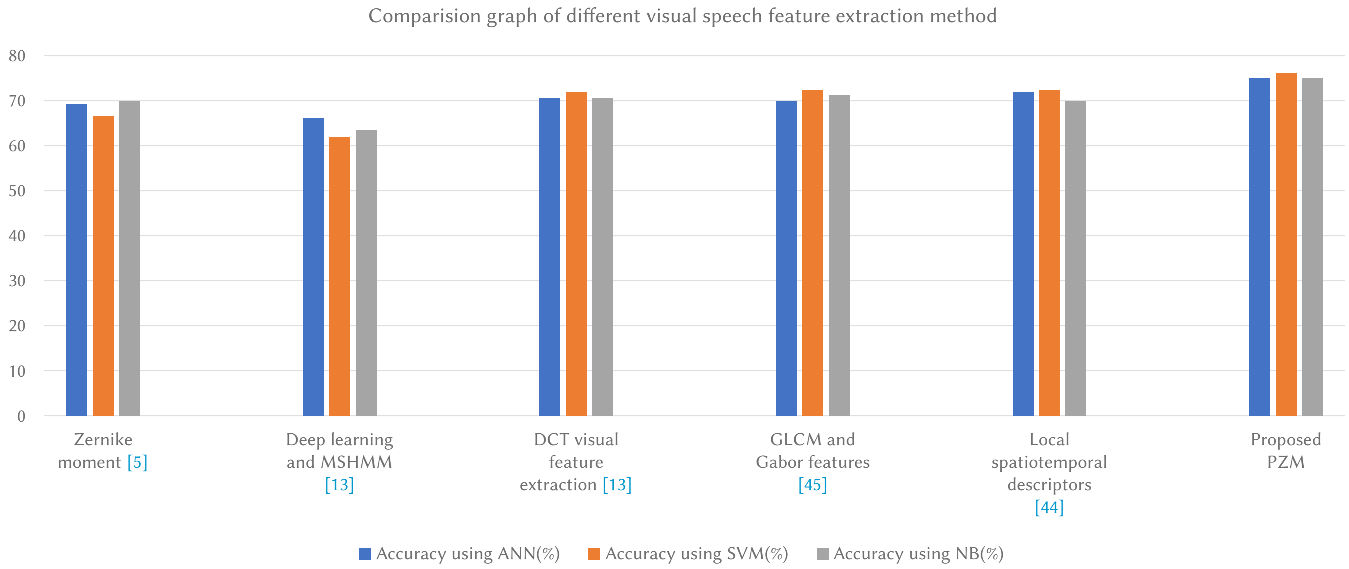


Fig. 4. Comparison of proposed visual speech recognition with existing method ('vVISWa' dataset).

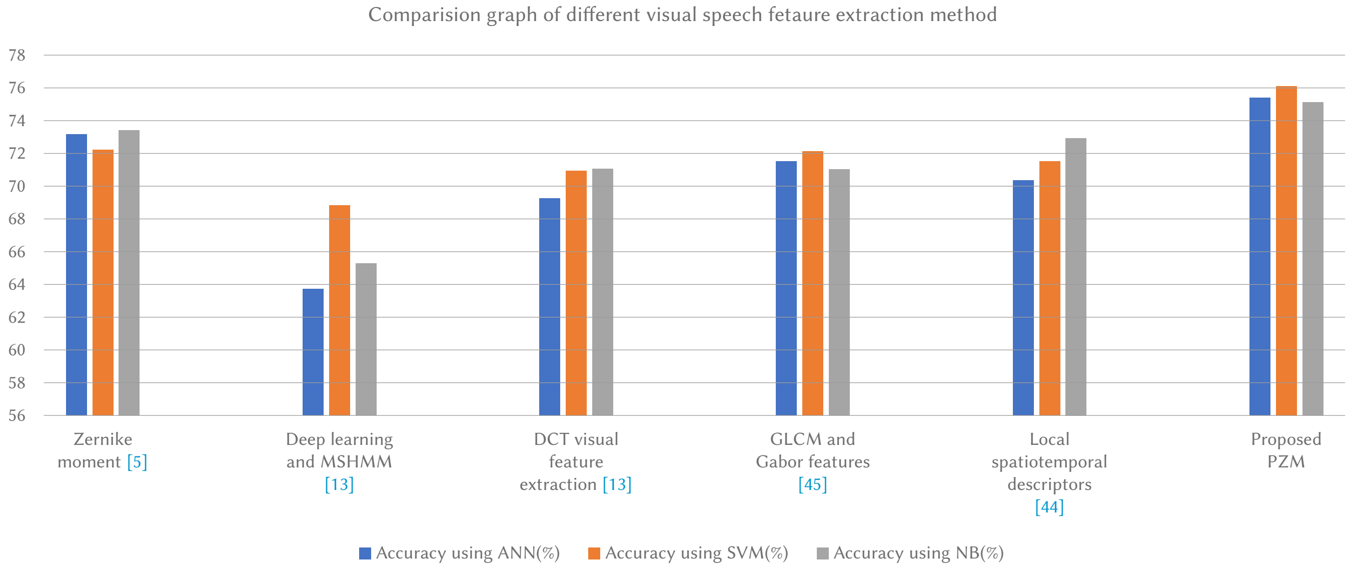


Fig. 5. Comparison of proposed visual speech recognition with existing method ('CUAVE' dataset).

for visual speech recognition. In this experiment, it has been observed that our proposed model using PZM gives better recognition accuracy than ZM for the both 'vVISWa' and 'CUAVE' English digit datasets, which is depicted in Fig. 4 and Fig. 5. Using PZM the system gives a better recognition rate because PZM has more feature dimensions and better feature representation capability, also PZM is less sensitive to noise than the ZM. We have also used local spatiotemporal descriptors [44] and GLCM and Gabor features [45] for lip reading in visual speech recognition. These are appearance-based features and capture the texture description of lip images. Appearance-based features consider all pixels within the ROI that are informative for speech recognition. However, the appearance-based features are sensitive to illumination, orientation variation, and position of the head [48]. Thus, these features are not very efficient for visual speech recognition. Shape-based features calculate the width and height of the lip contours of the speaker's lips. The proposed shape-based features extracted by PZM are illumination, rotation, translation, and scale-invariant. Therefore, in this research, we have introduced the shape-based feature extraction using PZM. The recognition rate of visual speech using GLCM and Gabor features are presented and compared in table XI. In audio speech recognition, when we select features based on feature selection algorithm it gives more accuracy than without any feature selection method. Using ANOVA, Kruskal-wallis and Friedman test statistical algorithm we have extracted the important features and model the system accordingly. The feature selection algorithm is important to remove the redundant features as well as to rank the significant features. All the individual classifiers select the feature subset using the feature selection method. In this paper, we have considered Zernike moment based visual speech recognition [5], deep learning and MSHMM [13], DCT visual feature extraction [43], local spatio-temporal descriptors [44], and GLCM and Gabor features [45] for comparison of visual speech recognition using 'vVISWa' and 'CUAVE' dataset and MFCC and HMM [19], MFCC and GMM [20], MFCC, SMRT and SVM [23] and optimal feature selector based on ABC-PSO [24] are taken for the comparison of audio speech recognition using our proposed model. Fig. 4 and Fig. 5 depict the comparison of the proposed visual speech recognition with existing methods for 'vVISWa' and 'CUAVE' dataset respectively. Fig. 6 and Fig. 7 represent the comparison graph of audio speech recognition of the proposed model and the existing models for both 'vVISWa'

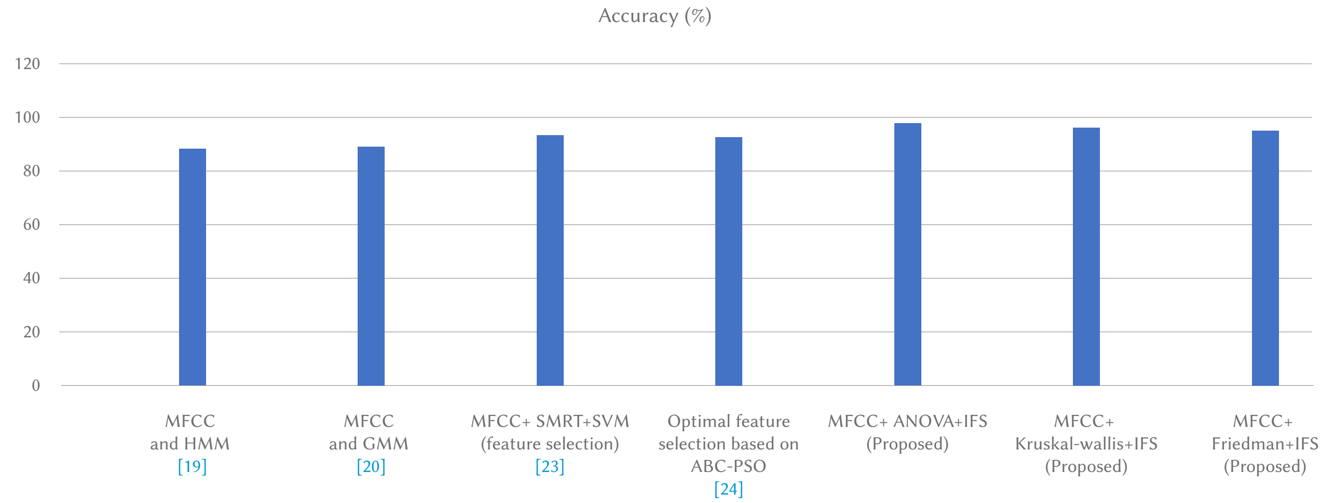


Fig. 7. Comparison of proposed audio speech recognition with existing method ('CUAVE'dataset).

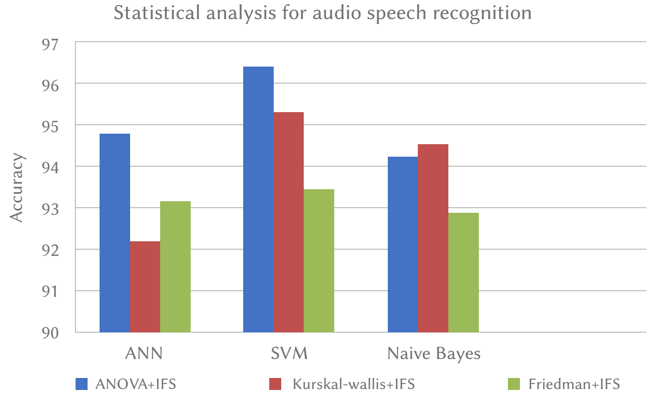


Fig. 6. Performance of different statistical analysis and classifiers for audio speech recognition ('vVISWa' dataset).

and 'CUAVE' dataset respectively. If both the systems are analysed individually, for only audio speech it gives a good recognition rate in a lab environment, but in a noisy environment, the audio signal may get corrupted, also the signal may degrade because of the different environment channels. In visual speech recognition, the face may not be properly detected all the time because of the camera and different lightning issues, also, a person's lip may not move properly all the time. But when we combine both the systems, it helps to overcome the shortcomings of individual recognition. The proposed system combines audio-visual recognition at the decision level and recognizes the speech based on audio-visual features.

From the experiment it has been shown that our proposed model gives better recognition accuracy than the existing methods. 'CUAVE' dataset provides better recognition accuracy than the 'vVISWa' dataset for both audio and visual speech.

VI. Conclusion and Future Work

In AV-ASR, the primary research on developing algorithms for the lip-reading, representation of visual features, and the integration of audio-visual information are the most promising areas. By watching the speaker's lip movement along with his voice can improve speech intelligibility especially in a noisy background and for hearing impaired people. Though it is an emerging field of research it still lacks proper visual articulations for visual speech recognition. Thus, the extraction of proper visual articulation attracts the interest of researchers in AV-ASR. Different types of lip-reading conditions provide very significant information regarding visual speech. This research has proposed shape-based visual speech features used for classification by the machine learning algorithms. The system includes two individual recognition: visual speech recognition and audio speech recognition. Visual speech recognition comprises of face detection, ROI detection, and lip tracking. A new visual feature extraction method using PZM has been proposed to track the lip movement. The PZM is an efficient orthogonal moment that describes a discriminating feature of an image or frame. In an audio speech, MFCC has been used and statistical algorithm along with IFS for selecting the significant features is proposed. The proposed method ranks the features based on their statistical significance and select features subset for the individual classifier. This paper compares the results using a feature selection method and without any feature selection method. After recognition combining the two modalities of audio-visual speech at the decision phase it gives the final outcome of AV-ASR. We use the threshold-based decision fusion and the threshold has been taken based on the average accuracy of individual recognition. The research can be extended in the future to develop a system using more specific audio-visual speech feature in a real-time environment. In the real-time, sometimes features may not be recognized properly because of noise, improper articulations. Thus, it is essential to capture more speaker-independent visual features.

References

[1] N. Moritz, K.Adiloglu, J. Anemuller, S. Goetze, B. Kollmeier, "Multi-Channel Speech Enhancement and Amplitude Modulation Analysis for Noise Robust Automatic Speech Recognition", Computer Speech & Language, vol. 46, pp. 558-573, 2017.

[2] D. Rudrapal, S. Das, S. Debbarma, N. Kar, N. Debbarma, "Voice Recognition and Authentication as a Proficient Biometric Tool and its Application in Online Exam for P.H People", International Journal of Computer Applications (0975 8887), vol. 39, no. 12, 2012.

[3] S. Singh and M. Yamini, "Voice based login authentication for Linux," 2013 International Conference on Recent Trends in Information Technology (ICRTIT), 2013, pp. 619-624, doi: 10.1109/ICRTIT.2013.6844272.

[4] Z. Saquib, N. Salam, R. Nair, N. Pandey, "Voiceprint Recognition Systems for Remote Authentication-A Survey", International Journal of Hybrid Information Technology, vol. 4, no. 2, 2011.

[5] P. Borde, A. Varpe, R. Manza, P. Yannawar, "Recognition of Isolated Words using Zernike and MFCC features for Audio Visual Speech Recognition", International Journal of Speech Technology, 2014, doi: 18.10.1007/s10772-014-9257-1.

[6] G. F. Meyer, J. B. Mulligan, S. M. Wuerger, "Continuous audiovisual digit

recognition using N-best decision fusion", Information Fusion, vol. 5, 2004.

[7] H. Marouf and K. Faez, "Zernike Moment-Based Feature Extraction For Facial Recognition Of Identical Twins", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), vol. 3, no. 6, 2013.

[8] A. Bhatia and E. Wolf, "On the Circle Polynomials of Zernike and Related Orthogonal Sets", Proceedings of the Cambridge Philosophical Society, vol. 50, no.1, pp. 40-48, 1954.

[9] S. B. Davis, P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-365, 1980.

[10] H. Ding, P. M. Feng, W. Chen, H Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis", MolBioSyst, vol. 10, no.8, pp. 22292235, 2014.

[11] A. Ganapathiraju, J. E. Hamakerand, J. Picone, "Applications of Support Vector Machines to Speech Recognition", IEEE Transactions on Signal Processing, vol. 52, no. 8, 2004.

[12] P. Borde, R. Manza, B. Gawali and P. Yannawar. " Article: vVISWa A Multilingual Multi-Pose Audio Visual Database for Robust Human Computer Interaction", International Journal of Computer Applications, vol. 137, no. 4, pp. 25-31, 2016.

[13] K. Noda, Y. Yamaguchi, K. Nakadai, H. Okuno, T. Ogata, "Audio-visual speech recognition using deep learning", Applied Intelligence, vol. 42, 2014, doi:10.1007/s10489-014-0629-7.

[14] G. Meyer, J. Mulligan, S. Wuerger, "Continuous audio-visual digit recognition using N-best Decision Fusion", Information Fusion, vol. 5, pp. 91-101, 2004.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg and A. W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech", Proceedings of the IEEE, vol. 91, no. 9, pp. 1306-1326, 2003, doi: 10.1109/JPROC.2003.817150.

[16] N. Dave, "A Lip Localization Based Visual Feature Extraction Method", An International Journal (ECIJ), vol. 4, no. 4, 2015.

[17] A. G. Chitu, L.J.M Rothkrantz, J. C. Wojdel, W. Pascal, "Comparison Between Different Feature Extraction Techniques for Audio-Visual Speech Recognition", Journal on Multimodal User Interfaces, vol. 1, no. 1, pp 720, 2007.

[18] S. Dupont and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition", IEEE Transacctions on Multimedia, Vol. 2, No. 3, 2000.

[19] T. S. Gunawan, A. A. M. Abushariah, M. A. M. Abushariah and O. Othman, "English Digits Recognition System Based on Hidden Markov Models," IEEE 978-1-4244-6235- 3/10/ , 2010.

[20] H. R. Goyal and S. G. Koolagudi, "Hindi Number Recognition using GMM," Global Journal of Computer Applications, vol. 63, no. 21, pp. 25-30, 2013.

[21] A. Jalalvand, F. Triefenbach, K. Demuynck, and J.-P. Marten, "Robust continuous digit recognition using Reservoir Computing," Computer Speech and Language , vol. 30, no. 1, pp. 135-158, 2015.

[22] S. Lokesh, P. M. Kumar, M. R. Devi, P. Parthasarathy, C. Gokulnath , "An Automatic Tamil Speech Recognition system by using Bidirectional Recurrent Neural Network with Self Organizing Map", Neural Computing and Applications , vol. 31, pp. 1521-1531, 2019, doi: https://doi.org/10.1007/s00521-018-3466-5.

[23] Mini P P, T. Thomas, R Gopikakumari, "Feature Vector Selection of Fusion of MFCC and SMRT Coefficients for SVM Classifier Based Speech Recognition System", 978-1-5386-6575-6 /18/$31.00 2018 IEEE.

[24] S. Mendiratta, N. Turk, D. Bansal, "Automatic Speech Recog- nition Using Optimal Selection of Features Based On Hybrid ABC-PSO", 2016 International Conference on Inventive Computation Technologies (ICICT), doi: 10.1109/INVENTIVE.2016.7824866.

[25] R. Lienhart, J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection", Intel Labs, Intel Corporation, Santa Clara, n.d. Web. 2014.

[26] E. Gregori. "Introduction To Computer Vision Using OpenCV. Embedded Vision Alliance", 2012 Embedded Systems Conference, 2012.

[27] P.I. Wilson, J. Fernandez, "Facial feature detection using haar classifiers", Texas A & M University, (2014).

[28] C. Teh, R. Chin, "On Image Analysis by the Method of Moments", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.10, no. 4, pp. 496-513, 1988.

[29] C. Singh,R. Upneja, "Accurate calculation of high order pseudo Zernike moments and their numerical stability", Digital Signal Processing, vol. 27, pp. 95-106, 2014.

[30] K. M. Hosny , "Accurate pseudo Zernike moment invariants for greylevel images", The Imaging Science, vol. 60, doi: 10.1179/1743131X11Y.0000000023.

[31] R. Mukundan, K.R. Ramakrishnan, "Moment Functions in Image Analysis Theory and Applications", World Scientific, Singapore (1998).

[32] G. Chandrashekar, F. Sahin, " A survey on feature selection methods", Computer Electrical Engineering, vol. 40, no. 1, pp. 628, 2014.

[33] B. Soni, S. Debnath, P.K. Das, "Text-dependent speaker verification using classical LBG, adaptive LBG and FCM vector quantization", International Journal of Speech Technology September, vol. 19, no. 3, pp. 525-536, 2016.

[34] M.A. Hossan, S. Memon, M.A. Gregory, "A Novel Approach for MFCC feature extraction", 4th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-5, 2010.

[35] N. Settouti, M.E.A. Bechar, M.A. Chikh, "Statistical comparisons of the top 10 algorithms in data mining for classification task", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4 no. 1, pp. 46-51, 2016.

[36] B. Niu, G. Huang, L. Zheng, X. Wang, F. Chen, Y. Zhang, T. Huang, "Prediction of substrate-enzyme-product interaction based on molecular descriptors and physicochemical properties", BioMed Research International, vol. 2013, article ID 674215, 2013.

[37] Y. Chan, R. P. Walmsley, "Learning and understanding the Kruskal Wallis one-way analysis of variance by ranks test for differences among three or more independent groups", Physical Therapy, vol. 77 no. 12, pp.1755-1761,1997 .

[38] D.W. Zimmerman, B.D. Zumbo, "Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks", Journal of Experimental Education, vol. 62, no. 1, pp. 75-86, 1993.

[39] J.D. Pujari, R. Yakkundimath, A.S. Byadgi, "SVM and ANN based classification of plant diseases using feature reduction technique", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 3, no. 7, pp. 6-14, 2016.

[40] W. Gevaert, G. Tsenov, V. Mladenov, "Neural Networks used for Speech Recognition", Journal of Automatic Control. Vol. 20, pp. 1-7, 2010.

[41] S. Russell, P. Norvig, "Artificial intelligence: a modern approach", 2nd edn. Prentice Hall, Englewood Cliffs. ISBN 978-0137903955, 2003.

[42] M. Islam, A. Roy, R. H. Laskar, "SVM-based robust image watermarking technique in LWT domain using different sub-bands," Neural Computing and Applications, vol. 32, pp. 1379-1403, 2020, doi: https://doi.org/10.1007/s00521-018-3647-2.

[43] A. Jain and G. N. Rathna, "Visual Speech Recognition for Isolated Digits Using Discrete Cosine Transform and Local Binary Patterns Features", 978-1-5090-5990-4/17/, 2017 IEEE.

[44] G. Zhao, M. Barnard, M. Pietikainen, "Lipreading with local spatiotemporal descriptors", IEEE Transactions on Multimedia, vol. 11, no. 7, 1254-1265, 2009.

[45] A. Kandagal, V. Udayashankara, "Visual Speech Recognition Based on Lip Movement for Indian Languages", 2017.

[46] N. Saleem, M. Khattak, E.Verdú, "On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 2, pp. 78-89, 2020, doi:10.9781/ijimai.2019.12.001.

[47] N. Saleem, M. Khattak, "Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp. 84-90, 2020, doi: 10.9781/ijimai.2019.06.001.

[48] M. Z. Ibrahim, D. J. Mulvaney, "Robust geometrical-based lip-reading using Hidden Markov models", Eurocon 2013, pp. 2011-2016.

[49] E. K. Patterson, S. Gurbuz, Z. Tüfekci, J.N. Gowdy, "CUAVE: A new audio-visual database for multmodal human-computer interface research", 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002, pp. II-2017-II-2020, doi: 10.1109/ICASSP.2002.5745028.

Saswati Debnath

Saswati Debnath received the B.Tech. degree in Computer Science and Engineering from Government Women Engineering College, Ajmer, Rajasthan Technical University India in 2013 and the M.Tech. degree in Computer Science and Engineering from National Institute of Technology, Silchar, Assam, India in 2015. She received Ph.D. from National Institute of Technology, Silchar, Assam, India in 2020.

Pinki Roy

Dr. Pinki Roy received the B.Tech. and M.Tech. degrees in Computer Science and Engineering from Dr. Babasaheb Ambedkar Technological University, Lonere, Maharastra, India, and the Ph.D. in Computer Science and Engineering from National Institute of Technology, Silchar, Assam,India. Currently, she is an assistant professor with National Institute of Technology, Silchar, Assam, India.

Neighborhood Structure-Based Model for Multilingual Arbitrarily-Oriented Text Localization in Images/Videos

H. T. Basavaraju[1]*, V.N. Manjunath Aradhya[1], D.S. Guru[2]

[1] Department of Computer Applications, JSS Science and Technology University (SJCE), Mysuru, Karnataka (India)

[2] Department of Studies in Computer Science, University of Mysore, Mysuru, Karnataka (India)

Received 14 April 2020 | Accepted 4 April 2021 | Published 4 May 2021

Abstract

The text matter in an image or a video provides more important clue and semantic information of the particular event in the actual situation. Text localization task stands an interesting and challenging research-oriented process in the zone of image processing due to irregular alignments, brightness, degradation, and complex-background. The multilingual textual information has different types of geometrical shapes and it makes further complex to locate the text information. In this work, an effective model is presented to locate the multilingual arbitrary oriented text. The proposed method developed a neighborhood structure model to locate the text region. Initially, the maxmin cluster is applied along with 3X3 sliding window to sharpen the text region. The neighborhood structure creates the boundary for every component using normal deviation calculated from the sharpened image. Finally, the double stroke structure model is employed to locate the accurate text region. The presented model is analyzed on five standard datasets such as NUS, arbitrarily oriented text, Hua's, MRRC and real-time video dataset with performance metrics such as recall, precision, and f-measure.

Keywords

Multilingual Text, Arbitrarily-Oriented, Maxmin Cluster, Neighborhood Structure, Double Stroke Structure.

DOI: 10.9781/ijimai.2021.05.003

I. Introduction

An image or a frame without textual information is tough to realize the precised situation in the real-life environment. Therefore, the text information present in a picture or a frame delivers the broad range of evidence of an incident. A person can easily realize the text information in a picture or a video. But, the computer could not understand the text as alike human. Hence, computer researchers follow the text extracting processes such as detection, localization, segmentation, and recognition. The text localization step plays a substantial role in the text understanding process for video indexing, video retrieval, video tracking, and video understanding. The main challenges encountered during the localization process are the complex background, shearing, low-resolution, illumination, embossed, night vision, alignment, variation in font size, color and style. This paper presents the neighborhood structure-based model to identify the text edges. The maxmin cluster is applied to discriminate the textual pixels from the unwanted pixels. Normal deviation helps to distinct the textual edges from other object edges in the neighborhood structure concept. Finally, the double stroke structure model is employed to identify the location of the real textual region.

II. Related Works

Ample of methods have been developed for text localization in horizontal directions, but a small quantity of approaches have been suggested to locate the multi-oriented textual information, and very less amount of models have been developed to locate the multilingual textual content [1]. Aradhya and Pavithra [2], [3], [4] worked on two-dimensional wavelet decomposition operation to identify the text information. Unar et al., [5] presented MSER, Sobel, and Canny edge detection algorithms to extract the text area. Dutta et al. [6] identify the actual scene text using entropy-based properties and histogram of oriented gradients. Jiang et al. [7] used an improved stroke feature transform, MSER and frequency tuned visual saliency to locate the textual information in natural scene images. Shivakumara et al. [8] worked on gradient and color properties to identify the location of the text region. He et al. [9] introduced a cascaded convolution text network (CCTN) to the coarse-to-fine text localization process. Gabor, wavelet, and k-means algorithms were employed by Pavithra and Aradhya [10], [11] to locate the actual textual area. Laplacian of Gaussian and full connected component concept is applied by Basavaraju et al. [12] to locate the textual content. Shekar et al. [13] proposed a text localization method using DWT and gradient difference model to obtain true text contents. Neumann and Matas [14] developed a text localization model using the sliding window concept, connected component, strokes and gradient concepts to locate the actual text blocks. A probabilistic model is introduced by Basavaraju et al. [15] to identify the actual text contents using hidden Markov random field, E-M algorithm. Xue et al. [16] presented a model to locate the textual contents in low-resolution pictures and videos using gradient values, low pass filter, and Bhattacharyya distance. Liu et al. [17] implemented a CTD concept to identify the curved text information using a transverse and longitudinal offset connection model. Li et

* Corresponding author.

E-mail address: basavaraju.com@gmail.com

al. [18] introduced a progressive scale expansion network (PSENet) to locate the text information. Xie et al. [19] introduced a supervised pyramid context network (SPCNET) to extract the region of the actual text space. Satwashil and Pawar [20] developed a hybrid technique to isolate the textual information from the scene pictures using character descriptor properties and SVM classifiers. Busta et al. [21] applied a trainable convolution neural network to locate the scene text region. Wu et al. [22] developed a strip-based text detection network (STDN) and a region proposal network to identify the location of the text region. Panda et al. [23] tuned the parameters of the MSER model for localizing the multilingual text present in scene images. The maximally stable extremal region (MSER) parameters are manually tuned to analyze the image dimension, text size and text region area. Villamizar et al. [24] developed a multi-scale fully convolutional and sequential network to segment the textual information. U shape network (UNet) identifies the text features from the several resolution images. Finally, the semantic text segmentation network helps to refine the semantic textual information. Zhang et al. [25] developed a fast dense residual network using fast residual dense blocks to recognize the character. With this literature knowledge, the present chapter discussed the segmentation process by applying a level set model with Gaussian mixture model and recognition process by implementing VGG-16 neural network. Ghoshal and Banerjee [26] implemented a model for segmenting and recognizing the character in scene images using canny edge technique, multi-layer perceptron and SVM.

Most of models were developed to identify the multilingual text based on certain combinations of features, but there is no a generalized model to identify the multilingual text. Hence, this paper introduces a neighborhood structure-based generalized model to locate the region of arbitrarily-oriented multilingual scene and graphical text present in images or videos.

III. Proposed Methodology

The presented method introduced the neighborhood pixel variance model to identify the location of the text candidates in images or videos. Primarily, the specified color image or video sequence (frame) is divided into R, G and B components. Later, the maxmin cluster concept is applied to enhance the textual space using three color values. Again the maxmin cluster is applied along with a 3X3 sliding window concept on the enhanced frame to sharpen the text edges. A flexible threshold is considered by calculating the normal deviation from the sharpened image. The neighborhood pixel variance model aids to produce the border lines for text instance of a picture or a frame based on normal deviation. Finally, the double stroke structure model is employed to determine the exact text space. Fig. 1 depicts the graphical representation of the presented algorithm.

A. Clustering of Color Information for Text Edge Enrichment

The maxmin cluster is applied to group the color pixel values for text region enhancement by suppressing the non-text region. The maxmin cluster extracts three color components such as R, G, and B bands from the specified color image or frame. For each RGB pixel, the maximum, minimum and middle intensity levels are separated. If a middle intensity level is close to a maximum intensity level then the respective pixel belongs to max-cluster. Similarly, if the middle intensity level is close to a minimum intensity level then that pixel is determined as min-cluster. Finally, the maxmin cluster model helps to enhance the input image or video frame. Further, the distance between the text space and non-textual space is increased by applying again the maxmin cluster concept along with the 3X3 sliding window on the enhanced frame. The maxmin cluster with the 3X3 sliding window concept results in the sharpened image is as shown in Fig. 3b.

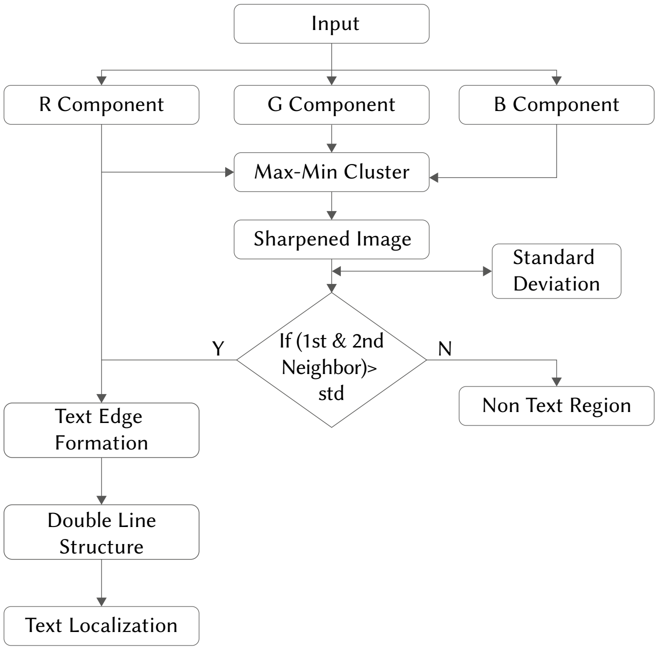


Fig. 1. Graphical representation of the proposed model.

B. Normal Deviation for Text Localization

Normal deviation (nd) is a geometric dimension and it can also be termed as standard deviation, which aids to evaluate the scattering of information based on mean value. The probable error is estimated by calculating the dissimilarity among each and every intensity levels. Therefore, the normal deviation is applied to compute the dissimilarity of every pixel in the sharpened frame. Primarily, the normal deviation calculates the mean score and then it computes the square root of deviation for all intensity levels from its mean value. The minimum normal deviation value denotes that entire intensity levels are closely spaced. The maximum normal deviation value denotes that entire intensity levels are distributed in a wide range. With this idea, the normal deviation(nd) is calculated from the sharpened image to identify the actual text edges. The calculation of normal deviation is represented in Equation (1).

$$nd = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})}{N-1}} \tag{1}$$

Where:

nd: Normal deviation.

N: The Size of data points.

X_i: Represent each pixel value.

$\overline{X}$: Depicts the mean value of all pixel values.

C. Neighborhood Pixel Variance Method

The neighborhood structure model helps us to originate the edge information for the textual region in the image or a frame. The presented model emphasizes the borderlines and suppresses the constant section. In the picture or frame sequences, an intensity level delivers substantial information about the space. If the pixel value stands nearly constant in the environment, then respective pixel is considered as area pixels. If the pixel levels have sudden variation, then corresponding pixel is determined as edge pixels. In the actual life situation, maximum of the textual candidates are terminated with its individual borderline as relate to non-textual space. Therefore, the neighborhood structure model has presented to obtain these edge intensity levels in the image or video sequences.

The possible textual region is obtained by differencing the pivot pixel from the 1st and 2nd neighborhood intensity levels. Initially, flexible threshold is measured by computing the normal deviation for the sharpened image. Later, the R component of the given input is considered to extract the potential text edges. For each pivot pixel in the R component, the first neighbor consists of eight picture elements and the second neighbor involves sixteen picture elements. The pivot picture element is differenced with every eight and sixteen picture elements. If the subtracted value is bigger than the flexible threshold, then the respective picture element is determined as a prominent text picture element, else the respective picture element is determined as a non-text picture element. Finally, the neighborhood structure model produces the edge information for text components in the specified picture or frame. Equation 2 is the representation of the potential text pixel extraction process. In the equation 2, f(x,y) represents the pivot pixel, if the difference between 1st neighborhood pixels, 2nd neighborhood pixels, and pivot pixel is larger than nd (normal deviation), then respective f(x,y) intensity level is replaced by 1, otherwise it is allocated by 0. This process continues for each and every intensity level to yield the edges for text candidates. The 1st and 2nd neighborhoods of a pivot picture element are illustrated in Fig. 2.

1	2	3
4	0	5
6	7	8

(a)

1	2	3	4	5
6				7
8		0		9
10				11
12	13	14	15	16

(b)

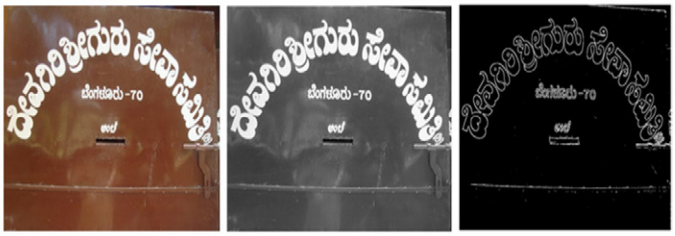
Fig. 2. (a) First neighborhood, (b) Second neighborhood.

$$f(x,y) = \begin{Bmatrix} if\ [f(x,y) - \{ I\&\,II\ Neighbors\} > \ nd],\ \ 1 \\ if\ [f(x,y) - \{ I\ \&\ II\ Neighbors\} < \ nd],\ \ 0 \end{Bmatrix} \tag{2}$$

Where:

nd: Normal deviation.

x and y: Space related coordinates.

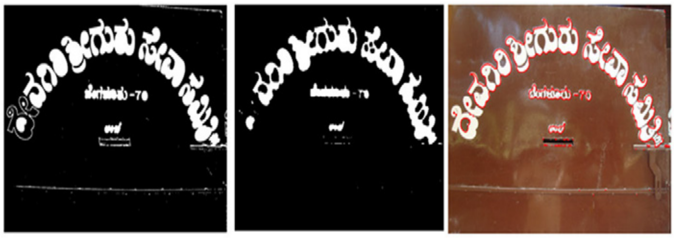


a. Input Image b. Gray Level c. Possible Text

Fig. 3. Extraction of possible text component process.

D. True Text Candidates

In real-life situation, the textual area seems with circular arbitrarily-oriented shape is depicted in Fig. 3c. According to the organization of textual components, it understands that the actual textual components can be obtained from the outcome of the neighborhood structure model. Therefore, the internal space of circular arbitrarily oriented region is filled by performing the morphological operation. In this specific direction, the specified RGB picture or frame is enhanced by employing the maxmin cluster concept and sharpened by the 3X3 sliding window. Consequently, the latent text information is identified by employing the innovative pixel variance model. Then true text components are obtained by applying the double-line structure approach [8]. If the initial and terminating points of the component are similar, then it is termed a double stroke structure model or arbitrary oriented circular shape. The neighborhood structure model draws the double stroke structure model of text parts. The internal holes of the arbitrary oriented shapes are occupied by applying morphological operation. Lastly, actual text components are recognized by taking the difference between the resultant of the neighborhood structure model represented in Fig. 3c and the filled region shown in Fig. 4a. Fig. 4 presents an extraction of the actual text components using the neighborhood structure model. Fig. 4c signifies the localized textual regions of the actual text components (i.e., Fig. 4b).



a. Holes Filled b. True Text Region c. Final Output

Fig. 4. Extraction of true text region and final output.

IV. Experimental Analysis

The presented model is verified on benchmark datasets like NUS dataset, Hua's dataset [27], the arbitrarily-directional dataset referred in [28], MRRC [29] and ICDAR 2013 video dataset. These datasets enfold all kinds of problems with multiple languages. The effectiveness of the presented model is calculated based on text lines. Precision (Eq. (4)), recall (Eq. (3)) and f-measure (Eq. (5)) evaluating factors are calculated to determine the efficiency of a presented method. The evaluating factors are computed by the following parameters. Real Textual Block (RTB) exhibits total amount of textual blocks in a picture or frame. Truly Located textual Block (TLB) depicts the textual space identified by the presented model. Fallaciously Located textual Block (FLB) refers non-textual space located by the presented model. The presented model employed the neighborhood pixel variance model using eight and sixteen neighbor pixels to generate the important textual spaces. The double-line structure helps us to identify the actual textual space from the outcome of the neighborhood pixel variance concept. The subsequent subsections depict the experimental analysis of five different datasets.

$$Recall(R) = \frac{TDB}{RTB} \tag{3}$$

$$Precision(P) = \frac{TDB}{TDB + FDB} \tag{4}$$

$$F - Measure(F) = \frac{2RP}{R + P} \tag{5}$$

A. Experimental Analysis on NUS Dataset

The NUS database contains 62 pictures. This database consists of straight textual line and curved textual line pictures with composite scenes and lighting effects. The presented model separates the text space from the non-text space. The outcome of the presented model on the NUS database is demonstrated in Fig. 5. Table I shows the qualified analysis of the presented model with a formerly existing method.

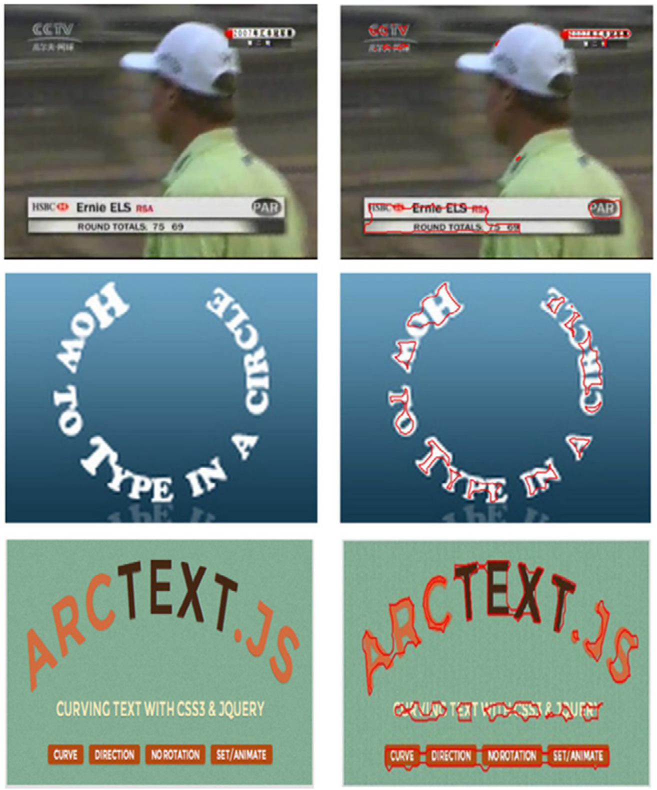


Fig. 5. Inputs and equivalent outcomes of NUS dataset.

TABLE I. Qualified Analysis of the Presented Model with an Earlier Existing Model on NUS Dataset

Methods	R	P	F
Shivakumara et al. [30]	85	84	82
Proposed method	91.77	81.32	85.14

B. Experimental Analysis on Hua's Dataset

Hua's database contains 45 pictures with straight textual lines. This database is gathered from sports and news programs. The proposed model has efficiently localized the straight line text region along with fewer false alarms in contrast variation. Fig. 6 represents the example outcomes of the presented model. Table II concludes that the presented model outperforms in recall and f-measure.

TABLE II. Qualified Analysis of the Presented Model with Earlier Existing Models on Hua's Data

Methods	R	P	F
Zhou et al. [31]	72	82	77
Wong and Chen [32]	51	75	61
Sharma et al. [33]	88	77	82
Fourier-RGB [34]	81	73	76
Lu et al. [35]	75	54	63
Bayesian [36]	87	85	85
Proposed method	91.85	80.17	85.64

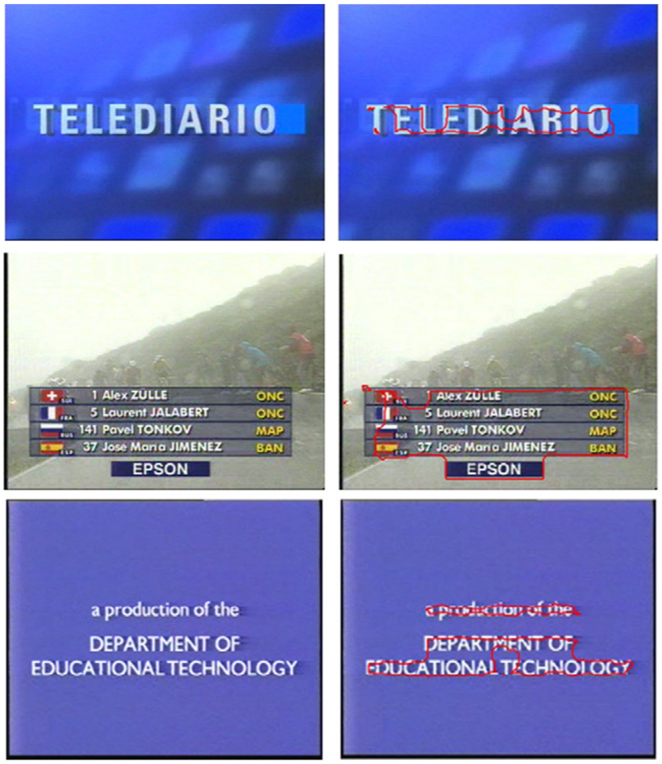


Fig. 6. Inputs and equivalent outcomes of Hua's dataset.

C. Experimental Analysis on Arbitrariness-Oriented Dataset

An arbitrariness directional database contains 142 pictures with blended textual lines, low contrast, composite background, and lighting effects. The presented model effectively localizes the text space along with few false alarms in composite backgrounds. Fig. 7 depicts the example outcomes of the presented model. Table III represents the quantified outcome of the presented model. The extracted outcome concludes that the presented model outperforms in recall and f-measure.

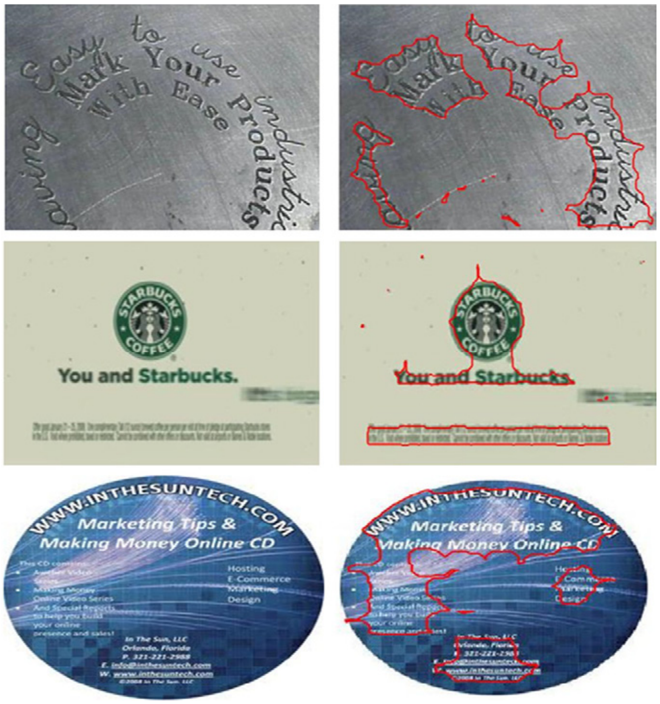


Fig. 7. Inputs and equivalent outcomes of arbitrariness directional dataset.

TABLE III. Qualified Analysis of the Presented Model with Earlier Existing Models on an Arbitrariness Directional Dataset

Methods	R	P	F
Zhou et al. [31]	41	60	48
Wong and Chen [32]	34	90	49
Sharma et al. [33]	73	88	79
Fourier-RGB [34]	52	68	58
Lu et al. [35]	47	54	50
Bayesian [36]	59	52	55
Proposed method	85.25	81.39	80.21

D. Experimental Analysis on MRRC-334 Dataset

MRRC-334 expands that Multi-script Robust Reading Competition. This database contains 167 training frames and 167 testing frames. A diversity of textual localization issues are enfolded in this database. The main issues are blended, lighting effect, artistic, night visualization, obstruction, scratch, shiny and deepness with multiple languages. The presented model effectively locates the textual space with less false positives. Fig. 8 and 9 depict an outcome of the presented model. Table IV represents the testing outcomes of the presented model on the MRRC database. The extracted outcome overtakes all the parameters.

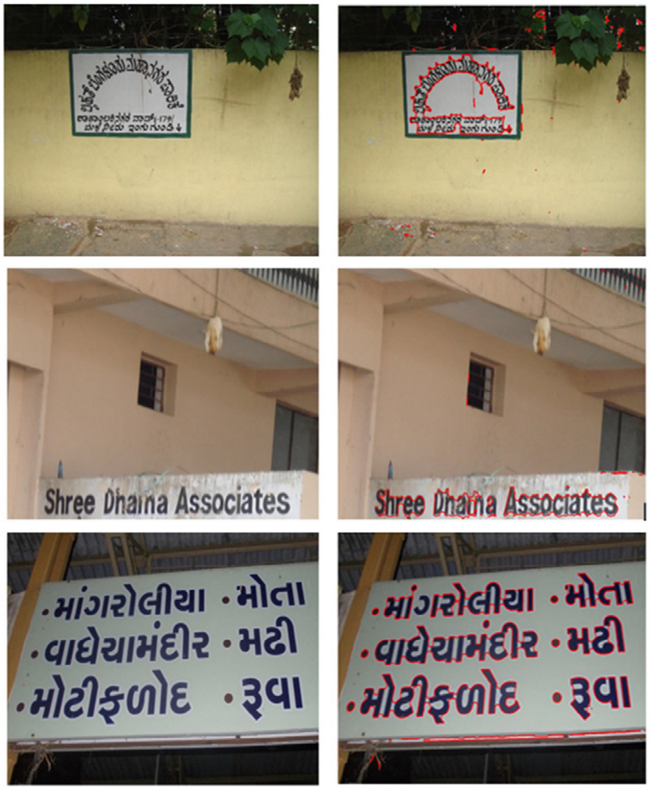


Fig. 8. Inputs and equivalent outcomes of MRRC training dataset.

TABLE IV. Qualified Analysis of the Presented Model with Earlier Existing Model on MRRC Dataset

Methods	R	P	F
Yin et al. [37]	64	42	51
Proposed method	75.37	74.85	71.93

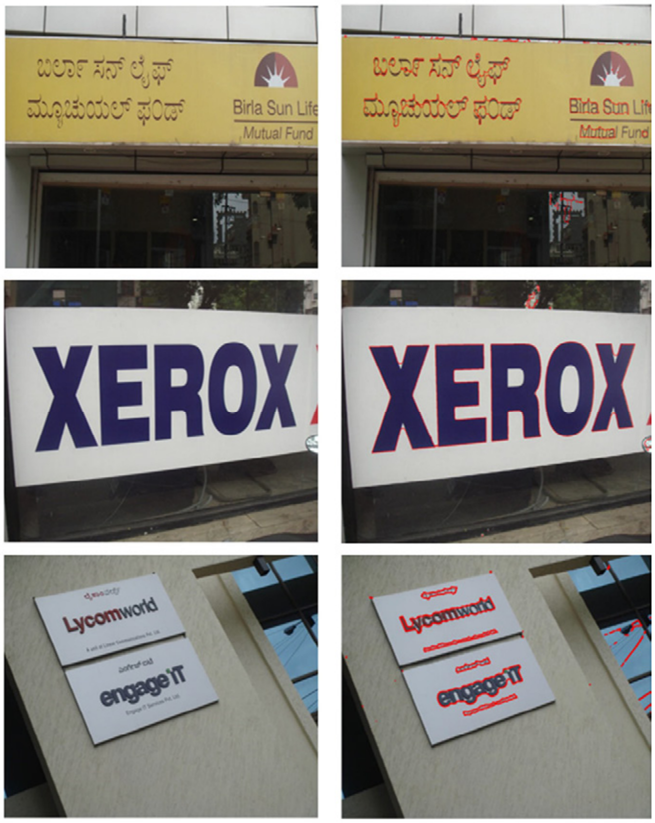


Fig. 9. Inputs and equivalent outcomes of MRRC testing dataset.

E. Experimental Results on ICDAR 2013 Video Dataset

The presented model has also conducted experimentation on real-time videos. This video dataset is collected from ICDAR 2013 dataset. This dataset contains the scene text along with unwanted noise. With this dataset is difficult to locate the actual text blocks due to low-resolution, distortion, occlusion, illumination effect and complex background. The proposed approach is employed on all video frames. The proposed model conducts the experimentation on the subset of frame sequences to compute the efficiency. The presented model effectively and efficiently identifies the text blocks from video with recall 79.34, precision 71.63 and f-measure 75.28. Fig. 10 shows the example outcomes of the presented model on real-time videos.

The lower link shows the investigational results of video dataset:

https://drive.google.com/drive/folders/1Oa_
BnddyLiyaiacZ_1gURbbraOIXXU5G

V. Conclusion

The proposed method developed an effective approach to locate the arbitrariness directional multilingual textual information in pictures or videos. This model is a broad-spectrum analyzed approach to identify the textual space. The maxmin cluster efficiently groups the color information to enhance the given frame. The sliding window concept increases the distance among the text space and the non-text space. The neighborhood pixel variance concept successfully locates the probable textual spaces and a double-line structure or closed arbitrary oriented circular shape effectively identifies the location of the actual textual space. The presented model conducts the evaluation on five standard datasets including video datasets by allowing all types of deviations. In future work, the current research work needs to be improved for the text segmentation process.

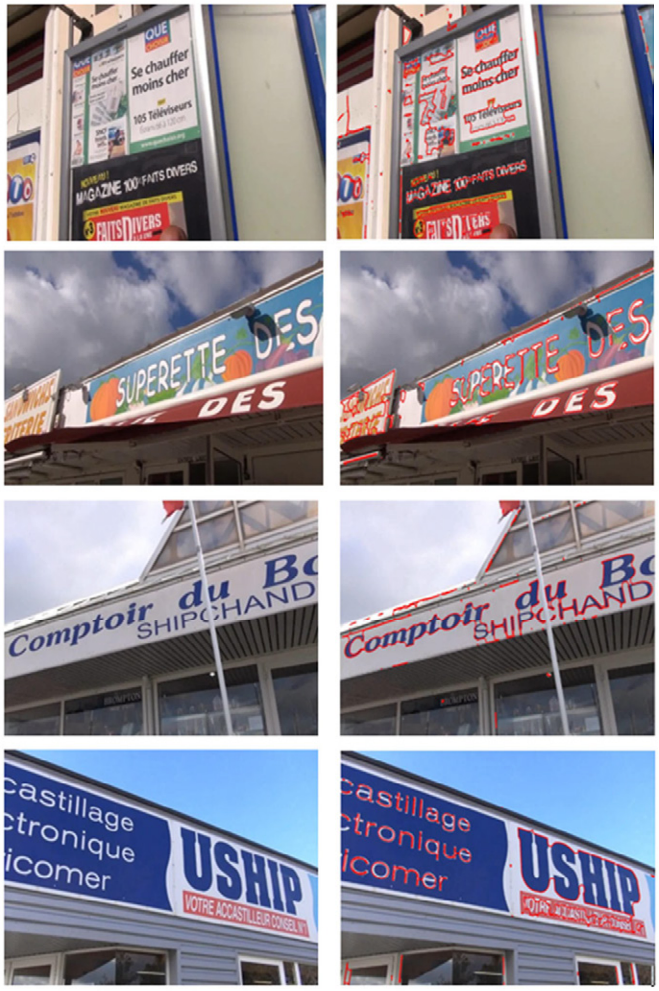


Fig. 10. Inputs and equivalent outcomes of video frames.

References

[1] V. N. Manjunath Aradhya, H. T. Basavaraju, and D. S. Guru, “Decade research on text detection in images/videos: a review,” Evolutionary Intelligence, 2019, pp. 1-27, https://doi.org/10.1007/s12065-019-00248-z.

[2] V.N. M. Aradhya, and M. S. Pavithra, “An application of LBF energy in image/video frame text detection,” 14th international conference on frontiers in handwriting recognition, 2014, pp.760–765.

[3] V. N. M. Aradhya, M. S. Pavithra and C. Naveena, “A robust multilingual text detection approach based on transforms and wavelet entropy,” Procedia Technology, 2012, pp. 232-237.

[4] V. N. M. Aradhya, M. S. Pavithra, and S. K. Niranjan, “An exploration of wavelet transform and level set method for text detection in images and video frames,” In Recent advances in intelligent informatics, 2014, pp. 419-426.

[5] S. Unar, A. H. Jalbani, M. M. Jawaid, M. Shaikh, and A. A. Chandio, “Artificial Urdu text detection and localization from individual video frames,” Mehran University research journal of engineering and technology, vol. 37, no. 2, 2018, pp. 429–438.

[6] K. Dutta, N. Das, M. Kundu, and M. Nasipuri, “Text localization in natural scene images using extreme learning machine,” In second international conference on advanced computational and communication paradigms (ICACCP), 2019, pp.1–6.

[7] M. Jiang, J. Cheng, M. Chen, and X. Ku, “An improved text localization method for natural scene images,” In journal of Physics: conference series, Vol. 960, No. 1, 2018, pp. 012027.

[8] P. Shivakumara, D. S. Guru, and H. T. Basavaraju, “Color and gradient features for text segmentation from video frames,” International conference on multimedia processing, communication and computing applications, 2013, pp.267–278.

[9] T. He, W. Huang, Y. Qiao, and J. Yao, “Accurate text localization in natural image with cascaded convolutional text network,” arXiv preprint arXiv:1603.09423,2016.

[10] M. S. Pavithra, and V. N. M. Aradhya, “A comprehensive of transforms, Gabor filter and k-means clustering for text detection in images and video,” Applied computing and informatics, 2014, pp. 1–15.

[11] V. N. M. Aradhya, and M. S. Pavithra, “An application of k-means clustering for improving video text detection,” In intelligent informatics, 2013, pp.41-47.

[12] H. T. Basavaraju, V. N. M. Aradhya, and D. S. Guru, “A novel arbitrary-oriented multilingual text detection in images/video,” In information and decision sciences, 2018, pp. 519–529.

[13] B. H. Shekar, M. L. Smitha, and P. Shivakumara, “Discrete wavelet transform and gradient difference based approach for text localization in videos,” In fifth international conference on signal and image processing, 2014, pp. 280–284.

[14] L. Neumann, and J. Matas, “Scene text localization and recognition with oriented stroke detection,” In Proceedings of the IEEE international conference on computer vision, 2013, pp. 97–104.

[15] H. T. Basavaraju, V. N. M. Aradhya, and D. S. Guru, “Text detection through hidden Markov random field and EM-algorithm,” In information systems design and intelligent applications, 2019, pp.19–29.

[16] M. Xue, P. Shivakumara, C. Zhang, T. Lu, and U. Pal, “Curved text detection in blurred/non-blurred video/scene images,” Multimedia tools and applications, 2019, pp. 1–25.

[17] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, “Curved scene text detection via transverse and longitudinal sequence connection,” Pattern Recognition, Vol. 90, 2019, pp. 337–345.

[18] X. Li, W. Wang, W. Hou, R. Z. Liu, T. Lu, and J. Yang, “Shape robust text detection with progressive scale expansion network,” arXiv preprint arXiv:1806.02559, 2018.

[19] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, “Scene text detection with supervised pyramid context network,” In proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 9038–9045.

[20] K. S. Satwashil, and V. R. Pawar “English text localization and recognition from natural scene image,” In international conference on intelligent computing and control systems (ICICCS), 2017, pp. 555–559.

[21] M. Busta, L. Neumann, and J. Matas, “Deep text spotter: An end-to-end trainable scene text localization and recognition framework,” In proceedings of the IEEE international conference on computer vision, 2017, pp. 2204–2212.

[22] D. Wu, R. Wang, P. Dai, Y. Zhang, and X. Cao, “Deep strip-based network with cascade learning for scene text localization,”. In 14th IAPR international conference on document analysis and recognition (ICDAR), Vol. 1, 2017, pp. 826–831.

[23] S. Panda, S. Ash, N. Chakraborty, A. F. Mollah, S. Basu, and R. Sarkar, “Parameter tuning in mser for text localization in multi-lingual camera-captured scene text images,” In computational intelligence in pattern recognition. Springer, 2020, pp. 999–1009.

[24] M. Villamizar, O. Can´evet, and J. M. Odobez, “Multi-scale sequential network for semantic text segmentation and localization,” in Pattern Recognition Letters, Vol. 129, Elsevier, 2020, pp. 63–69.

[25] Z. Zhang, Z. Tang, Y. Wang, J. Qin, H. Zhang, and S. Yan, “Fast dense residual network: Enhancing global dense feature flow for text recognition,” in arXiv preprint arXiv:2001.09021.

[26] R. Ghoshal and A. Banerjee, “Svm and mlp based segmentation and recognition of text from scene images through an effective binarization scheme,” In computational intelligence in pattern recognition. Springer, 2020, pp. 237–246.

[27] X. S. Hua, L. Wenyin, and H. J. Zhang, “An automatic performance evaluation protocol for video text detection algorithms,” IEEE Trans CSVT, 2004, pp. 498–507.

[28] C. Lu, C. Wang, and R. Dai, “Text detection in images based on unsupervised classification of edge based features”, In: Proceedings. ICDAR, 2005, pp. 610–614.

[29] Multi-script robust reading competition. http://mile.ee.iisc.ernet.in/mrrc/index.html.

[30] P. Shivakumara, H. T. Basavaraju, D. S. Guru, and C. L. Tan, “Detection of curved text in video: quadtree based method,” In: 12th international conference on document analysis and recognition (ICDAR), 2013, pp.

594–598.

[31] J. Zhou, L. Xu, B. Xiao, and R. Dai, "A robust system for text extraction in video," In: Proceedings of ICMV, 2007, pp. 119–124.

[32] E. K. Wong, and M. Chen, "A new robust algorithm for video text extraction," Pattern Recognition, 2003, pp. 1397–1406.

[33] N Sharma, P. Shivakumara, U. Pal, M Blumenstein, and C. L. Tan, "New method for arbitrarily oriented text detection in video," In: Proceedings of DAS, 2012, pp. 74–78.

[34] P. Shivakumara, T. Q. Phan, and C. L. Tan, "New Fourier-statistical features in RGB space for video text detection," IEEE Transaction on CSVT, 2010, pp. 1520–1532.

[35] C. Lu, C. Wang, and R. Dai, "Text detection in images based on unsupervised classification of edge based features," In: Proceedings of ICDAR, 2005, pp. 610–614.

[36] P. Shivakumara, R. P. Sreedhar, T. Q. Phan, S. Lu, and C. L. Tan, "Multi-oriented video scene text detection through Bayesian classification and boundary growing," IEEE Trans. CSVT, 2012, pp. 1227–235.

[37] X. C. Yin, X. Yin, K. Huang, and H. W. Hao, "Robust text detection in natural scene images", IEEETrans. PAMI 36, 2014, pp. 970–983.

H T Basavaraju

H T Basavaraju received the B.Sc. and MCA degrees in Computer Science from the University of Mysore, Mysuru, Karnataka, India in 2009 and 2012 respectively. He worked as a research officer at AIISH from 2012 to 2013. He received the Ph.D. degree at the Department of Computer Applications, Sri Jayachamarajendra College of Engineering, Visveswaraya Technological University, Karnataka, India. His research interest includes Document image analysis, Computer vision, Speech processing, Machine learning, Deep Learning, Natural language processing and Artificial Intelligence.

V.N. Manjunath Aradhya

Dr. V.N. Manjunath Aradhya is currently working as an Associate Professor & Head in the Dept. of Computer Applications, JSS Science and Technology University, Mysuru. He received the M.S. and Ph.D. degrees in Computer Science from the University of Mysore, Mysuru, India, in 2004 and 2007 respectively. He is a recipient of "Young Indian Research Scientist" from the Italian Ministry of Education, University and Research, Italy during 2009-2010. An awardee of "Young Scientist" from the Department of Science and Technology (DST) in 2009 under FAST TRACK SCHEME. Recently awarded prestigious ARP (Award for Research Publications) by Vision Group on Science and Technology (VGST), Govt. of Karnataka for the year 2016-17. His professional recognition includes as a Technical Editor for Journal of Convergence Information Technology (JCIT), Editor Board in Journal of Intelligent Systems, reviewer for IEEE Trans on System, Man, and Cybernetics - PART B, Pattern Recognition (PR), and Pattern Recognition Letters (PRL). His research interest includes Pattern Recognition and Image Processing, Speech Processing, Document Image Analysis, Computer Vision, Machine Intelligence, Applications of Linear Algebra for the Solution of Engineering Problems, Biclustering of Gene Expression Data and Web Data Analysis and Understanding.

D. S. Guru

Prof. D. S. Guru received his B.Sc, M.Sc and Ph.D. degrees in Computer Science and Technology from the University of Mysore, Mysuru, India in 1991, 1993 and 2000 respectively. He is currently a Professor in the Department of Studies in Computer Science, University of Mysore, India. He was a fellow of BOYSCAST and a visiting research scientist at Michigan State University. He has authored 65 journals and 225 peer-reviewed conference papers at international and national levels. His area of research interest covers image retrieval, text mining, machine learning, object recognition, shape analysis, sign language recognition, biometrics, and symbolic data analysis.

Performance and Convergence Analysis of Modified C-Means Using Jeffreys-Divergence for Clustering

Ayan Seal[1,2]*, Aditya Karlekar[3], Ondrej Krejcar[2,4], Enrique Herrera-Viedma[5,6]

[1] PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, 482005 (India)
[2] Center for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, 50003 (Czech Republic)
[3] Hitkarini College of Engineering and Technology, Jabalpur, 482005 (India)
[4] Malaysia Japan International Institute of Technology, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur (Malaysia)
[5] Department Computer Science and Artificial Intelligence, University of Granada, 18071 Granada (Spain)
[6] Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, 21589 (Saudi Arabia)

Received 27 September 2020 | Accepted 20 March 2021 | Published 29 April 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

The size of data that we generate every day across the globe is undoubtedly astonishing due to the growth of the Internet of Things. So, it is a common practice to unravel important hidden facts and understand the massive data using clustering techniques. However, non- linear relations, which are essentially unexplored when compared to linear correlations, are more widespread within data that is high throughput. Often, non-linear links can model a large amount of data in a more precise fashion and highlight critical trends and patterns. Moreover, selecting an appropriate measure of similarity is a well-known issue since many years when it comes to data clustering. In this work, a non-Euclidean similarity measure is proposed, which relies on non-linear Jeffreys-divergence (JS). We subsequently develop c- means using the proposed JS (J-c-means). The various properties of the JS and J-c-means are discussed. All the analyses were carried out on a few real-life and synthetic databases. The obtained outcomes show that J-c-means outperforms some cutting-edge c-means algorithms empirically.

Keywords

C-mean, Clustering, Convergence, Jeffreys-Divergence, Jeffreys-Similarity Measure.

DOI: 10.9781/ijimai.2021.04.009

I. Introduction

Machine learning considers clustering to be an important issue. It is normally used to reveal some existing hereditary structure by analyzing a set of data items or patterns. The aim of clustering is to split data into groups so that data in the same groups are similar and data items in different groups are not capable of comparison in the same sense. Clustering is the subject of active research for varying areas, including marketing [1], biology [2], libraries [3], insurance [4], city planning [5], and earthquake studies [6]. Common clustering algorithms include Gaussian Mixture models [7], hierarchical clustering [8], Hidden Markov models [9], self-organizing maps [10], and c- means clustering [11]. Hierarchical clustering constructs a multi- level hierarchy of groups by making a tree, which is known as a cluster tree. Gaussian mixture model forms groups, which would be considered as a mixture of multivariate normal density components. The self-organizing map takes the help of neural networks for learning the topology and data structure in the form of distribution. Hidden Markov models use observed data for recovering the sequence of states.

The performance of a clustering algorithm always relies upon data items or their features, choice of the initial cluster centers, similarity measures, objective function, and clustering algorithms [12]–[14]. In this study, the *c*-means algorithm is implemented on synthetic and real-life databases, so everything is similar except the similarity measure. In other words, the use of different similarity measures is studied because the selection of proper similarity measures is an important issue in clustering and it helps to find the cluster structure in data [15] properly. However, Euclidean distance is one of the widely accepted similarity measures even though a large number of researches are going on around the world to introduce non-linearity in similarity measures for data clustering [15], [16]. In recent times, Euclidean distance in *c*-means is replaced using different non-linear metrics. From this, some do not obey triangle inequality property [17]–[21]. The objective of instigating non-linearity is to detect a more accurate boundary between two clusters. A. Banerjee et al. initiated general Bregman divergence as a distance metric in the *c*-means to augment its effectiveness [17]. This method in reality unified the divergence measures, for which the first moment was used as cluster

* Corresponding author.
E-mail address: ayanseal30@ieee.org

representative ensuring a gradual depreciating of the objective function in the iterative relocation technique. The interested reader can go through [12], [22]–[27] to know the use of various divergence-based similarity measures in clustering.

II. Clustering

This section presents the formal definition of clustering. A concise overview of conventional *c*-means is also discussed, given that the performance comparison is made between the conventional *c*-means and the modified one.

A. Basic Principle

The method of dividing *n*´dimensional *m* data-points or their features, $A[= a_1, a_2, ..., a_n]$, in R^n into 'c' groups of homogeneous data-points, $G[= (G_1, G_2, ..., G_c)]$ to increase association strength within the cluster, is known as clustering. However, association strength will be low or weak between different clusters. Then

$G_i \neq \phi \quad for \quad i = 1,\dots,c,$
$G_i \cap G_j = \phi \quad for \quad i = 1,\dots,c; j = 1,\dots,c \quad and \ i \neq j,$
$\cup_{i=1}^{c} G_i = G$

B. The C-Means Algorithm

It is certainly a well-known clustering technique because it is easy to implement. Sometimes, it is applied in the pre-processing step to finding the knowledge by analyzing data [28]. It partitions data into 'c' distinct groups by reducing the entire intra-cluster variance, beginning with an arbitrarily selected group of the centroid from each group. Each centroid should effectively denote the central location of a group. The ideal value of 'c' leads to the highest separation (distance) and is an unknown priori. It has to be approximated from the database itself. The c-means intends to reduce total intra-cluster variance, or, the squared error function, *E*, which could be computed using Eq. (1).

$$E = \sum_{j=1}^{m}\sum_{i=1}^{c} |\mathrm{a}_j - \mathrm{g}_i|^2 \tag{1}$$

where $|a_j - g_i|^2$ is a similarity measure between the cluster center, g_i and a data-object, a_j.

The *c*-means algorithm consists of the given steps:

Step 1: Select 'c' initial cluster centers $g_1, g_2, ..., g_c$ arbitrarily from the m data-points $A[= a_1, a_2, ..., a_n]$.

Step 2: Designate data-point a_j , *j*=1, 2, ..., m to cluster center g_i, *i* ∈ 1, 2, ..., *c* iff $||a_j - g_i|| \leq ||a_j - g_k||$, *k* = 1, 2, ..., *c*, &*i* ≠ *c*. Ties are broken randomly.

Step 3: Find new cluster centers $\mathrm{g}_1^+, \mathrm{g}_2^+, \dots, \mathrm{g}_c^+$, by Eq. (2).

$$\mathrm{g}_i^+ = \frac{1}{m_i}\sum_{\mathrm{a}_j \in G_i} \mathrm{a}_j \, , i = 1,2,\dots,c \tag{2}$$

where m_i is the count of data-objects in cluster G_i.

Step 4: If $\mathrm{g}_i^+ = g_i \forall i = 1, 2, ..., c$ then stop. If not, go to Step 2.

Note that if Step 4 does not terminate then the algorithm executes for a predetermined fixed number of epochs.

This work focuses to introduce JS, which is inherited from the concept of Jeffreys-divergence [29]. Several characteristics of this similarity measure are studied. The entire experiment set is executed on some synthetic and real-life benchmark databases. These simulation outcomes show that c-means utilizing JS performs better than a traditional c-means algorithm and along with *c*-means with various other divergences in certain situations. Our assertion is confirmed through a statistical analysis of the results obtained.

III. Jeffreys-Similarity Measure (JS) and its Properties

The definition of JS and its properties are discussed in this section.

Definition 3.1. Let J_n be a set of all positive definite matrices of size $n \times n$ and Jeffreys-divergence is a similarity measure defined over J_n, which could be computed by Eq. (3).

$$\partial_{Jeffreys}(P, Q) = (P - Q)(\log(|P|) - log(|Q|)) \tag{3}$$

where |*P*|=determinant of *P*.

Consider a real positive vector a $=(a_1, a_2, ..., a_n) \in \mathbb{R}_+^n$. Let us define a one-to-one function $\psi: \mathbb{R}_+^n \longrightarrow J_n$ such that ψ (a) = diag$(a_1, a_2, ..., a_n)$. The definition of JS is as follows:

Definition 3.2. The JS function $d_{Jeffreys}: \mathbb{R}_+^n \times \mathbb{R}_+^n \longrightarrow \mathbb{R}_+ \cup \{0\}$ between any two $a, b \in \mathbb{R}_+^n$ is defined by applying Eq. (4).

$$d_{Jeffreys}(\mathrm{a,b}) = \partial_{Jeffreys}(\psi(\mathrm{a}), \psi(\mathrm{b})) \tag{4}$$

The JS measure, $d_{Jeffreys}$, is well-stated because ψ is a one-to-one function by definition. Some of the following properties are stated here as $\partial_{Jeffreys}$ divergence is defined on J_n.

Proposition 3.1. $d_{Jeffreys}(a, b) = d_{Jeffreys}(b, a)$

Proof: $d_{Jeffreys}(a, b) = \partial_{Jeffreys}(\psi(a), \psi(b)) = \partial_{Jeffreys}(\psi(b), \psi(a)) = d_{Jeffreys}(b, a)$

Proposition 3.2.

$d_{Jeffreys}(a, b) \geq 0$ and $d_{Jeffreys}(a, b) = 0$ iff $a = b$

Proof: $d_{Jeffreys}(a, b) = \partial_{Jeffreys}(\psi(a), \psi(b)) \geq 0$ and $d_{Jeffreys}(a, b) = 0$ iff $\partial_{Jeffreys}(\psi(a), \psi(b)) = 0$ iff $\psi(a) = \psi(b)$ iff $a = b$

So, $d_{Jeffreys}$ is a similarity measure on $\mathbb{R}_+^n$, which could be thought as $d_{Jeffreys}(a, b) = \sum_{i=1}^{n} \partial_{Jeffreys}(a_i, b_i)$. Now, its time to investigate some of the properties of JS.

Theorem 3.1. The JS is not a Bregman divergence.

Proof: If JS was a Bregman divergence $d_{Jeffreys}(a, b)$ would have been strictly convex in *a*. However, our objective is to prove that $d_{Jeffreys}(a, b)$ is not convex in *a*. We know that the JS, $d_{Jeffreys}$, could also be expressed by Eq. (5).

$$d_{Jeffreys}(\mathrm{a,b}) = \sum_{i=1}^{n}(a_i - b_i)(\log(a_i) - \log(b_i)) \tag{5}$$

The expression below can be acquired if the derivative of both sides of Eq. (5) is taken with respect to a_i. $\frac{\partial d_{Jeffreys}}{da_i} = 1 - \frac{b_i}{a_i} + \log(a_i) - \log(b_i)$ $\frac{\partial^2 d_{Jeffreys}}{da_i da_j} = 0$ when $i \neq j$ otherwise,

$$\frac{\partial^2 d_{Jeffreys}}{\partial a_i^2} = \frac{b_i}{a_i^2} + \frac{1}{a_i}$$

We have, $\frac{\partial^2 d_{Jeffreys}}{\partial a_i^2} < 0$ for the values in the range of {-∞, -1} ∪ {0, 1}. So, $d_{Jeffreys}(a, b)$ is not convex in *a*. So, it is demonstrated that JS measure is not a Bregman divergence.

Theorem 3.2. $d_{Jeffreys}(x \circ a, x \circ a) = xd_{Jeffreys}(a, b)$ for $x \in \mathbb{R}_+^n$, where $x \circ a$ depicts the Hadamord product between *a* and *x*.

Proof: It is known that $(x \circ a) = (x_1 a_1, x_2 a_2, ..., x_n a_n)$. So,

$\delta_{Jeffreys}(x_i a_i, x_i b_i) = (x_i a_i - x_i b_i)(\log(x_i a_i) - \log(x_i b_i)) = x_i (a_i - b_i)$ $(\log x_i + \log a_i - \log x_i - \log b_i) = x_i (a_i - b_i)(\log a_i - \log b_i)$
$\sum_{i=1}^{n} \delta_{Jeffreys}(x_i a_i, x_i b_i) = \sum_{i=1}^{n} x_i \, \delta(a_i, b_i)$ implying
$d_{Jeffreys}(x \circ a, x \circ b) = x \, d_{Jeffreys}(a, b)$

Theorem 3.3. JS is f-divergence.

Proof: If a divergence expression can be made through the following

$\phi(t) = a \, \phi(\frac{\mathrm{b}}{\mathrm{a}})$, *where* $t = \frac{\mathrm{b}}{\mathrm{a}}$

then that divergence is known as f-divergence. The JS between $a \in \mathbb{R}_+^n$ and $b \in \mathbb{R}_+^n$ is given by

$d_{Jeffreys}$(a,q) $=\sum_{i=1}^{n}(a_i - b_i)(\log(a_i) - \log(b_i))$

putting $t_i = \frac{b_i}{a_i}$

$d_{Jeffreys}$(a,b) $=\sum_{i=1}^{n}(a_i - b_i t_i)(\log(a_i) - \log(a_i t_i))$

$=\sum_{i=1}^{n} a_i (1 - t_i)(\log(a_i) - \log(a_i) - \log(t_i))$

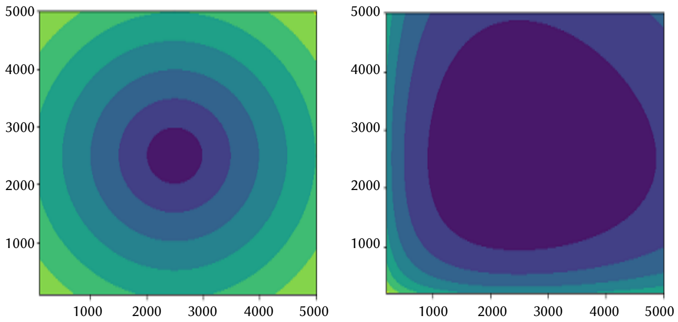
$=\sum_{i=1}^{n} a_i (1 - t_i)(-\log(t_i))$

$=\sum_{i=1}^{n} x_i (1 - t_i)(\log(\frac{1}{t_i}))$

$\sum_{i=1}^{n}\phi(t) = \sum_{i=1}^{n} x_i\, \phi(\frac{b_i}{a_i})$

Since, $d_{Jeffreys}$ (a, b) can be expressed as $\sum_{i=1}^{n} x_i\, \phi(\frac{b_i}{a_i})$. Thus, JS is f-divergence.

Remark 3.1: We may consider another imperative facet of JS. Fig. 1 portrays the contour plot of the norm-balls in $\mathbb{R}^2$ everywhere over the point (5000,5000) for Euclidean distance (Fig. 1a) and JS (Fig. 1b). We can also observe from Fig. 1 that the norm-ball of Euclidean distance is similar to concentric circles, on the other hand, JS is similar to some extent to askew ovals. It is further evident from Fig. 1b that contour lines confine together as they come near the origin i.e. (0,0). Thus, we conclude that the J-divergence between two points is greater when they come in the vicinity of the origin and it reduces when their distance from the origin increases. While on the contrary, the Euclidean distance within two points remains constant regardless of their location. For instance, the J-divergence and the Euclidean distance between (3,3) and (5,5) are 2.043 and 2.82 respectively and for points (1003,1003) and (1005,1005) they are 0.0079 and 2.82 respectively. At times, the attribute in question might prove beneficial in situations where the clusters have varying sizes and densities.



a) Contour plot of Euclidean distance

b) Contour plot of JS

Fig. 1. Contour plot of norm ball for the Euclidean distance and JS.

IV. Proposed Method

A. The C-Means with JS

Consider a given set of vector, $A = \{a_1, a_2, ..., a_m\}$, in $\mathbb{R}_+^n$. Our objective is to divide A into 'c' disjoint groups where, the value of 'c' could be any value between 2 and 'm'. This problem can be formalized using the following form.

M : *minimize* $\psi(W, G) = \sum_{j=1}^{m}\sum_{i=1}^{c} w_{ij}\, d_{Jeffrey}(a_j, g_i)$, subject to the constraints

$$\sum_{i=1}^{c} w_{ij} = 1 \tag{6a}$$

$$w_{ij} \in \{1,0\} \forall j \in \{1,\dots,m\}, \forall i \in \{1,\dots,c\} \tag{6b}$$

$$G = \{g_1, g_2, \dots, g_c\}, g_i \in \mathbb{R}_+^n, \forall i \in \{1,\dots,c\} \tag{6c}$$

Two following heuristics steps are given in order to solve M.

Initialization:

The 'c' number of vectors have to pick randomly from A and called them as cluster centers, which are denoted as

$$G^{(0)} = \{g_1^{(0)}, g_2^{(0)}, \dots, g_c^{(0)}\}$$

Iterative Steps:

- Set $W^{(z+1)} = argmin_W\, \psi(W, G^{(z)})$ subject to constraints 6a and 6b are satisfied. In other words, each vector a_i is assigned to its nearest cluster center.
- Set $G^{(z+1)} = argmin_G\, \psi(W^{(z+1)}, G)$ subject to constraint 6c is satisfied.
- Set $z = z + 1$ until convergence.

Criterion for stopping:

We cease iteration in cases where the cost function reduces experiences alteration i.e.

$\psi(W^{(z+1)}, G^{(z)}) = \psi(W^{(z)}, G^{(z)})$ or $\psi(W^{(z+1)}, G^{(z+1)}) = \psi(W^{(z+1)}, G^{(z)})$. An informal program code of J-c-means is given in algorithm 1.

```
Algorithm 1 J-c-means([A] m × n, c)
Input: a set of vector, $A = \{a_1, a_2, ..., a_m\}, a_i \in \mathbb{R}^n$.
Output: a partition, $M = \{A_1, A_2, ..., A_c\}$, of A together with the
    centroids $g_1, g_2, ..., g_c$ of each cluster.
Initialization: select $g_1, g_2, ..., g_c$ in A at random
while terminating condition has not been met do
    for $i = 1$ to $c$ do
        $A_i \leftarrow 0$
    end for
    for $j = 1$ to $m$ do //updating the class membership of the vectors
        $\omega(a_j) \leftarrow argmin_{i\in\{1,2,\dots,c\}}\, d_{Jeffreys}(a_j, g_i)$
        $A_{\omega(a_j)} \leftarrow A_{\omega(a_j)} \cup \{a_j\}$
    end for
    for $i = 1$ to $c$ do //updating centroids
        $m_i \leftarrow \sum_{j=1}^{n} 1\,(a_j \in A_i)$
        $g_i \leftarrow \frac{1}{m_i}\sum_{j=1}^{n} a_j\, 1\,(a_j \in A_i)$
    end for
    return $M, g_1, g_2, ..., g_c$
end while
```

B. Convergence of J-C-Means Algorithm

Theorem 3.1. The J-*c*-means monotonically decreases the inertia $\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{c} w_{ij}\, d_{Jeffreys}(a_j, g_i)$

Proof: Let $\phi(A^u) = \frac{1}{m}\sum_{i=1}^{c}\sum_{j=1}^{m} d_{Jeffreys}(a_j, g_i)$, where $A^{(u)}$ is the recent group $A_1^{(u)}, ..., A_c^{(u)}$ with the centre of the clusters $g_1^{(u)}, ..., g_c^{(u)}$and assignation function $\omega^{(u)}$, then $\phi(A^u) \geq \sum_{i=1}^{c}\sum_{a_j\in A_i^{(u)}} d_{Jeffreys}(a_j, g^{(u)}_{\omega^{(u+1)}_{(a_j)}})$ because $\omega(a_j)$ minimizes the quantity $d_{Jeffreys}(a_j, g_i)$ over all $i\in\{1,..., c\}$.

$\phi(A^u) \geq \sum_{i=1}^{c}\sum_{a_j\in A_i^{(u)}} d_{Jeffreys}(a_j, g_i^{(u+1)})$ because $g_i^{(u+1)}$ minimizes the quantity $d_{Jeffreys}(a_j, g_i)$ over all $a_j \in A_i$.

Therefore, $\phi(A^{(u)}) \geq (A^{(u+1)})$.

Corollary: The J-*c*-means stops after a finite amount of time.

There are only finite number of partitions $\binom{m}{c}$. Thus, the sequence $\phi(A^{(u)})_{u\in N}$ has a finite number of values i.e. there exist u such that $\phi(A^{(u+1)}) = \phi(A^{(u)})$.

Remark 4.1: The above corollary does not say anything about how fast the J-c-means converges. There is an exponential bound $\binom{m}{c}$. The time required for the above mentioned algorithm to converge depends on the initialization. However, some heuristic can be found in the literature.

V. Experiments

A. Database Description

All the experiments are performed on some synthetic databases: 2_blobs, 3_blobs, 5_blobs, and 10_blobs and real-word databases: Iris, Glass, Cleveland, Bank Note Authentication, Appendicitis, Breast Cancer Wisconsin, and Mammography. These real-world databases are acquired from the Keel Repository [30] and UCI Machine Learning Repository [31].

B. Cluster Validity Index

The fundamental question that requires to be responded to in clustering is: how good a clustering technique is. The concept of goodness is quantified by validity indexes. The notion of these indexes may be explained mathematically. We may consider *c*-partitions namely, $A_1, A_2, ..., A_c$ of A, found by a clustering technique and the valuations of their respective validity indexes are $Z_1, Z_2, ..., Z_c$. The $Z_{h1} \geq Z_{h2} \geq ... \geq Z_{hc}$ will represent that $A_{h1} \uparrow A_{h2} \uparrow ... \uparrow A_{hc}$, for a particular permutation $h1$, $h2$, ..., hc of {1, 2, ..., c}, where $A_i \uparrow A_j$ depicts that partition A_i is a better clustering than A_j [32]. Validity indexes can be categorized into two sets namely, internal validity index and external validity index. Two external validation indexes namely, Normalized Mutual Information (NMI) [33] and Adjusted Rand Index (ARI) [34] are considered in this work to measure the performance of the c-means algorithms by varying distance metrics. NMI will typically be utilized as an index that can compare the performance of two data-point groups. Meanwhile, ARI is seen as an index for cluster validation. Both metrics show the mismatch in terms of two data clustering of an allotted arrangement of data points. The highest value (1) and the lowest value (0) indicate no mismatch and complete mismatch respectively. Both metrics use the ground truth to compute the efficiency of a clustering algorithm. Three internal evaluation schemes, for example, the Silhouette index (SI) [35], Dunn index (DI) [32], and Davies Boulden Index (DBI) [32] are further employed in this research to explore the cohesiveness of the obtained clusters. These indexes estimate the similarity between a data point with the corresponding group called cohesion and disunion between different groups known as separation. The domain of SI lies within −1 and +1, in which a greater value illustrates that the data point is excellently suited with its corresponding cluster and weakly paired to neighboring clusters. A higher DI and lower DBI demonstrate a more favorable grouping.

C. Computational Protocols

Five sets of experiments were performed on the aforementioned databases through *c*-means-E: *c*-means with Euclidean distance [36], *c*-means-S: *c*-means with S-distance [37], *c*-means-W: Weighting in *c*-means [38], c-means-M: Minkowski weighted *c*-means [33], and *c*-means-P: the proposed c-means. Performance comparison: We consider the same arbitrarily selected centroids for all the algorithms while calculating ARI, NMI, SI, DI, and DBI values to make results consistent. The performance of a clustering algorithm does not rely on the better extraction of inceptive set centroids. Nevertheless, it relies upon the clustering technique. The exact methodology is administered tenfold on each database. Then Wilcoxon's rank-sum is executed to determine whether two dependent data-points from populations have the exact distribution on the acquired values of ARI, NMI, SI, DI, and DBI using the above-mentioned methods.

VI. Results and Discussion

Fig. 2 shows the clustering results. Table I shows the mean ARI, NMI, SI, DI, and DBI values obtained by the methods presented in section V-C on synthetic and real-life databases. However, the first two i.e. ARI and NMI are external clustering validity indexes for which actual class labels are required to match with the predicted class labels. database 2_blobs consists of two clusters having the same density and same size. However, one is close to the origin and the other is away from the origin. It is evident from Table I that the suggested c-means-P on 2_blobs defeats other algorithms mentioned in section V-C because nearly all of the ARI and NMI values are close to the greatest value i.e. 1. Moreover, c-means-P returns a higher expected value of ARI and NMI values over other algorithms, which depicts the efficiency of c-means-P. The proposed c-means-P outperforms due to askew oval figures of contour norm-balls of the J-divergence as considered in Remark 3.1. The proposed method also works well for the databases 3_blobs to 5_blobs, which contain clusters having the same size and same density. However, some noise is introduced to them. Still, the performance of the proposed method is good as J-divergence is invariant to the Hadamard product. The performance of all the methods on some real-life databases is noted in Table I. These outcomes depict that the proposed method c-means-P is the best among all the methods discussed in this study. The values of three internal clustering evaluation indexes namely, SI, DI, and DBI for the same databases are included in Table I. Although, actual class labels are not required in this case. The received results further validate the efficiency of the c-means-P over other methods discussed in section V-C due to the values obtained by c-means-P approach nearer to ideal values in comparison to values generated by methods other than the proposed one. The non-parametric Wilcoxon's rank-sum is also performed for comparing c-means-P over other methods presented in section V-C using the p-values achieved from ARI, NMI, SI, DI, and DBI. Table II reports the estimated p-values. We can very well observe that the generated outcomes advice that we discard the null hypothesis for a 5% level of significance. It may be proposed that substantial proof is presented using data available with us to comment that c-means-P algorithm surpasses other methods discussed in section V-C.

VII. Conclusion

In this work, a similarity measure on $\mathbb{R}^n_+$ is presented based on Jeffreys-divergence. Different JS properties are also elaborated. The conventional *c*-means algorithm is altered, where Euclidean distance is substituted with the similarity measure introduced. A theoretical evaluation of the JS and *c*-means was also conducted by outlining the convergence proof. Research on complexity metrics promises to be an area of research with potential when it comes to field clustering. It should be explored in future work. We focused on the evaluation of multiple database properties to find information. This can be used to design proper clustering algorithms. JS can be used for the Fuzzy *c*-means type algorithm.

Acknowledgment

This work is partially supported by the project "Prediction of diseases through computer assisted diagnosis system using images captured by minimally-invasive and non-invasive modalities", Computer Science and Engineering, PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur India (under ID: SPARC-MHRD-231). This work is also partially supported by the project Grant Agency of Excellence, Univer-sity of Hradec Kralove, Faculty of Informatics and Management, Czech Republic (under ID: UHK-FIM-GE-2204-2021).

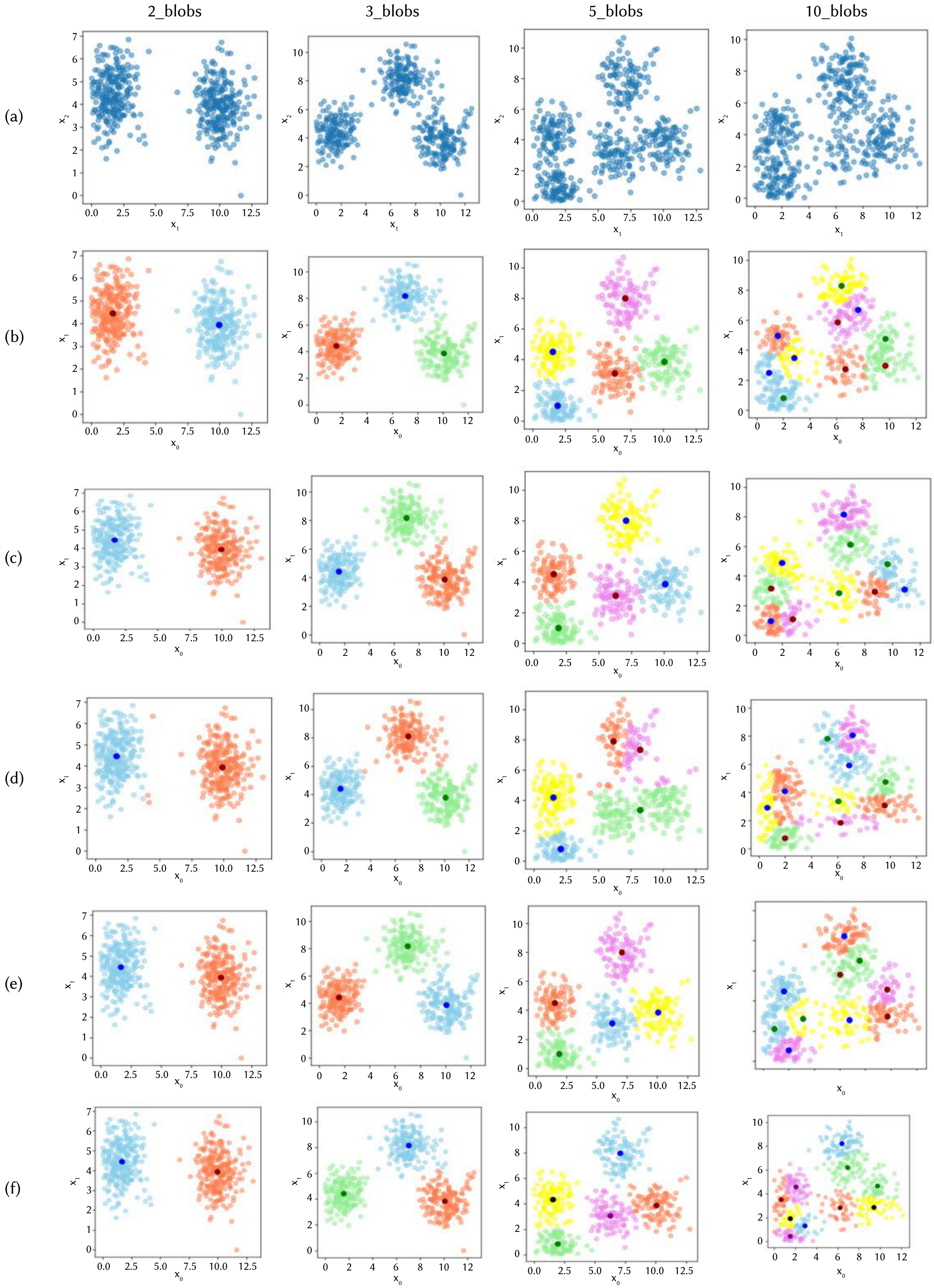


Fig. 2. (a): Original structure of #_blobs. Result of clustering corresponding #_blobs with (b): *c*-means-E, (c): *c*-means-M (d): *c*-means-S (e): *c*-means-W and (f): *c*-means-P.

TABLE I.: The Values of ARI, NMI, SI, DI and DBI for Synthetic and Real-life Databases

	Database	*c*-means-E	*c*-means-S	*c*-means-W	*c*-means-M	*c*-means-P
ARI	2_blobs	**1.0000000**	0.9760961	**1.0000000**	**1.0000000**	**1.0000000**
	3_blobs	**0.9820360**	0.9582248	0.9760904	0.9701813	**0.9820360**
	5_blobs	0.8986230	0.6279446	0.6663589	0.8941665	**0.9028355**
	10_blobs	0.430297	0.3859427	0.4502888	0.4185428	**0.4686642**
	Iris	**1.0000000**	**1.0000000**	**1.0000000**	0.9600667	**1.0000000**
	Glass	0.3008098	0.6361220	0.5476595	0.6276935	**0.7149283**
	Cleveland	0.0162569	**0.3369418**	0.0465415	0.0505246	0.1416707
	Bank Note Authentication	0.0485381	0.0404252	0.0485381	0.0488371	**0.0491855**
	Appendicitis	0.4843330	0.4320631	0.4843330	0.4978654	**0.5360417**
	Breast Cancer Wisconsin	0.4914245	0.5286179	0.4914245	0.4914245	**0.5666393**
	Mammography	0.0905026	0.0930185	0.0905026	0.0821258	**0.1275994**
NMI	2_blobs	**1.0000000**	0.9530566	**1.0000000**	**1.0000000**	**1.0000000**
	3_blobs	0.9541820	0.9391844	0.9362472	0.9503597	**0.9666142**
	5_blobs	**0.8883229**	0.7297619	0.7692242	0.8801728	**0.8883229**
	10_blobs	0.6232038	0.5858008	0.6035275	0.6132328	**0.6336725**
	Iris	**1.0000000**	**1.0000000**	**1.0000000**	0.9404430	**1.0000000**
	Glass	0.5075728	0.6577571	0.6441501	0.6832619	**0.7325871**
	Cleveland	0.0183458	0.1175260	0.0375054	0.0472017	**0.3864150**
	Bank Note Authentication	0.0303241	0.0245671	0.0303241	0.0312593	**0.0327895**
	Appendicitis	0.3999936	0.3690075	0.3809936	0.4048908	**0.4401108**
	Breast Cancer Wisconsin	0.4671655	0.4863613	0.4671655	0.4671655	**0.5163683**
	Mammography	0.0846832	0.0846832	0.0846832	0.0846832	**0.1298267**
SI	2_blobs	**0.7949160**	0.7874309	**0.7949160**	**0.7949160**	**0.7949160**
	3_blobs	**0.6932280**	0.6861834	**0.6932280**	0.6916559	**0.6932280**
	5_blobs	0.5711057	0.4126689	0.4446950	0.5759132	**0.5759364**
	10_blobs	0.3645875	0.2902406	0.3306109	0.3527441	**0.3857648**
	Iris	**0.5824192**	**0.5824192**	**0.5824192**	0.5818419	**0.5824192**
	Glass	0.2909336	0.1899170	0.3491109	0.2386990	**0.3928576**
	Cleveland	0.2076061	-0.026441	0.2657142	0.2390776	**0.2808949**
	Bank Note Authentication	0.4308310	0.4293403	0.4308310	0.4310046	**0.4310995**
	Appendicitis	0.4137615	0.4127611	0.4136630	0.4086627	**0.4137615**
	Breast Cancer Wisconsin	**0.6972643**	0.6741518	0.6910678	**0.6972643**	**0.6972643**
	Mammography	0.1243098	**0.5419065**	**0.5419065**	**0.5419065**	**0.5419065**
DI	2_blobs	**1.9040153**	1.3735754	1.3735754	**1.9040153**	**1.9040153**
	3_blobs	1.7047981	1.6202096	1.7047981	1.7047981	**1.7663088**
	5_blobs	**1.2592171**	0.6447496	0.6709407	**1.2592171**	**1.2592171**
	10_blobs	0.9048197	0.4620280	0.8099586	0.8140677	**1.2805434**
	Iris	**2.0197395**	**2.0197395**	**2.0197395**	1.9596349	**2.0197395**
	Glass	0.4010836	0.2986482	0.4644115	0.5325171	**0.6286726**
	Cleveland	0.5363801	0.5889773	0.5371508	0.5005224	**0.6011315**
	Bank Note Authentication	**1.5469099**	1.5013920	**1.5469099**	**1.5469099**	**1.5469099**
	Appendicitis	**1.0011285**	**1.0011285**	**1.0011285**	1.0017089	**1.0011285**
	Breast Cancer Wisconsin	**1.3494101**	1.1806848	**1.3494101**	1.3005589	**1.3494101**
	Mammography	**1.3974134**	**1.3974134**	**1.3974134**	1.1162343	**1.3974134**
DBI	2_blobs	**0.144604**	**0.144604**	**0.144604**	0.147063	**0.144604**
	3_blobs	0.157400	0.159603	**0.1565898**	**0.1565898**	**0.1565898**
	5_blobs	0.348582	**0.122678**	0.2365076	0.123528	0.1236736
	10_blobs	0.1068881	0.121510	0.162423	0.109133	**0.10423103**
	Iris	**0.167358**	0.16801707	**0.167358**	0.167373	**0.167358**
	Glass	0.532517	0.398066	0.46441156	0.271434	**0.2093515**
	Cleveland	0.383664	2.071026	0.33105801	0.4016002	**0.320763**
	Bank Note Authentication	0.4371350	0.436876	0.43713506	0.439077	**0.436666**
	Appendicitis	**0.516156**	0.5261876	**0.516156**	0.516380	**0.516156**
	Breast Cancer Wisconsin	0.268049	0.257680	**0.2522018**	**0.2522018**	**0.2522018**
	Mammography	**0.311799**	0.860389	0.34720557	**0.311799**	**0.311799**

TABLE II. P-Values Generated from ARI, NMI, SI, DI and DBI for Wilcoxon's Rank-sum Test for Comparing J-C-Means With other Algorithms

	Database	*c*-means-E	*c*-means-S	*c*-means-W	*c*-means-M
ARI	2_blobs	0.0010	1.5938E-06	0.0010	0.0010
	3_blobs	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
	5_blobs	5.3477E-06	6.1582E-06	4.0402E-06	4.7314E-06
	10_blobs	0.0099	0.0059	0.0099	0.0485
	Iris	0.0128	0.0088	4.0402E-06	4.0167E-04
	Glass	0.0046	0.01038	0.0017	0.0046
	Cleveland	1.4851E-04	1.8267E-04	0.0022	0.0211
	Bank Note Authentication	0.02547	6.0243E-06	4.5506E-06	0.01485
	Appendicitis	0.0325	6.0243E-06	0.0165	0.03681
	Breast Cancer Wisconsin	3.2899E-06	3.2899E-06	4.7314E-06	3.2899E-06
	Mammography	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
NMI	2_blobs	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
	3_blobs	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
	5_blobs	5.3477E-06	6.1582E-06	4.0402E-06	4.7314E-06
	10_blobs	0.04698	1.8165E-04	0.04097	0.04272
	Iris	0.0146	0.03812	5.7206E-06	4.0167E-04
	Glass	0.0013	0.0036	0.0013	0.0013
	Cleveland	1.4851E-04	1.8267E-04	0.0017	0.0058
	Bank Note Authentication	1.5938E-06	6.0243E-06	1.5938E-06	0.04339
	Appendicitis	0.0325	6.0243E-06	0.0125	0.0125
	Breast Cancer Wisconsin	1.5938E-06	1.5938E-06	1.5938E-06	2.4282E-06
	Mammography	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
SI	2_blobs	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
	3_blobs	0.00586	1.5938E-06	0.0332	0.0039
	5_blobs	6.1582E-06	6.1582E-06	0.1058	4.7314E-06
	10_blobs	0.01855	1.8165E-04	0.0211	0.0211
	Iris	0.0474	0.008812	0.0131	0.0469
	Glass	0.0451	1.8165E-04	0.0451	0.0451
	Cleveland	1.4851E-04	1.8267E-04	0.0204	0.04725
	Bank Note Authentication	2.4282E-06	0.04429	1.5938E-06	4.7682E-06
	Appendicitis	1.5938E-06	0.0010	0.03681	0.0165
	Breast Cancer Wisconsin	0.0010	2.1650E-06	2.1650E-06	2.1650E-06
	Mammography	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
DI	2_blobs	0.0010	0.0010	0.0010	0.0010
	3_blobs	0.0215	0.0215	0.0215	0.0215
	5_blobs	0.0014	0.045	0.0089	0.0078
	10_blobs	0.0339	0.0339	0.0339	0.07539
	Iris	0.02891	0.02891	0.02891	0.02891
	Glass	0.0339	0.0339	0.0339	0.0339
	Cleveland	0.03438	0.03438	0.03438	0.03438
	Bank Note Authentication	0.0010	0.0010	0.0010	0.0010
	Appendicitis	0.02547	0.0075	0.0125	0.0056
	Breast Cancer Wisconsin	0.0020	0.0020	0.0020	0.0020
	Mammography	0.0020	0.0020	0.0020	0.0020
DBI	2_blobs	1.5938E-05	1.5938E-06	1.5938E-06	1.5938E-06
	3_blobs	0.0486	1.5938E-06	0.0332	0.0039
	5_blobs	5.3477E-06	6.1582E-06	4.0402E-06	4.7314E-06
	10_blobs	0.0450	5.8006E-06	1.8267E-06	5.7729E-06
	Iris	0.0015	0.0321	9.6624E-06	2.5597E-06
	Glass	0.04722	1.8165E-06	0.01523	0.04772
	Cleveland	1.4851E-06	1.8267E-06	0.03845	0.0199
	Bank Note Authentication	2.4282E-06	0.04429	1.5938E-06	4.7682E-06
	Appendicitis	1.5938E-06	0.0014	0.03681	0.0013
	Breast Cancer Wisconsin	3.2899E-06	3.2899E-06	3.2899E-06	3.2899E-06
	Mammography	3.2899E-06	3.2899E-06	1.5938E-06	1.5938E-06

References

[1] H. Wattimanela, U. Pasaribu, S. Indratno, A. Puspito, "Eartquakes clustering based on the magnitude and the depths in molluca province," in *AIP Conference Proceedings*, vol. 1692, 2015, p. 020021, AIP Publishing.

[2] J. Yang, J. Cao, R. He, L. Zhang, "A unified clustering approach for identifying functional zones in suburban and urban areas," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WK- SHPS)*, April 2018, pp. 94–99.

[3] T. Pitchayaviwat, "A study on clustering customer sugges tion on online social media about insurance services by using text mining techniques," in *2016 Management and Innovation Technology International Conference (MITicon)*, Oct 2016, pp. MIT–148–MIT–151.

[4] R. Suresh, I. Anand, B. Vianesh, H. R. Mohammed, "Study of clustering algorithms for library management system," in *2018 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, March 2018, pp. 221–224.

[5] A. Naik, D. Reddy, P. K. Jana, "A novel clustering algorithm for biological data," in *2011 Second International Conference on Emerging Applications of Information Technology*, Feb 2011, pp. 249–252.

[6] K. C. Gull, A. B. Angadi, C. G. Seema, S. G. Kanakaraddi, "A clustering technique to rise up the marketing tactics by looking out the key users taking facebook as a case study," in *2014 IEEE International Advance Computing Conference (IACC)*, Feb 2014, pp. 579–585.

[7] J. Li, A. Nehorai, "Gaussian mixture learning via adaptive hierarchical clustering," *Signal Processing*, vol. 150, pp. 116–121, 2018.

[8] R. Abe, S. Miyamoto, Y. Endo, Y. Hamasuna, "Hierarchical clustering algorithms with automatic estimation of the number of clusters," in *17th World Congress of International Fuzzy Systems Association*, 2017.

[9] S. Ghassempour, F. Girosi, A. Maeder, "Clustering multivariate time series using hidden markov models," *International Journal of Environmental Research and Public Health*, vol. 11, pp. 2741–2763, 2014.

[10] M. Pacella, A. Grieco, M. Blaco, "On the use of self-organizing map for text clustering in engineering change process analysis: A case study," *Computational Intelligence and Neuroscience*, p. 11, 2016.

[11] V. Schellekens, L. Jacques, "Quantized compressive k-means," *IEEE Signal Processing Letters*, vol. 25, no. 8, 2018.

[12] K. K. Sharma, A. Seal, "Spectral embedded generalized mean based k-nearest neighbors clustering with s-distance," *Expert Systems with Applications*, p. 114326, 2020.

[13] K. K. Sharma, A. Seal, "Outlier-robust multi-view clustering for uncertain data," *Knowledge-Based Systems*, vol. 211, p. 106567, 2021.

[14] K. K. Sharma, A. Seal, "Multi-view spectral clustering for uncertain objects," *Information Sciences*, vol. 547, pp. 723–745, 2021.

[15] L. Bottou, Y. Bengio, "Convergence properties of the k-means algorithms," in *Advances in neural information processing systems*, 1995, pp. 585–592.

[16] A. Karlekar, A. Seal, O. Krejcar, C. Gonzalo-Martin, "Fuzzy k-means using non-linear s-distance," *IEEE Access*, vol. 7, pp. 55121–55131, 2019.

[17] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, "Clustering with bregman divergences," *Journal of machine learning research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[18] S. Chakraborty, S. Das, "k- means clustering with a new divergence-based distance metric: Convergence and performance analysis," *Pattern Recognition Letters*, vol. 100, pp. 67–73, 2017.

[19] L. Legrand, E. Grivel, "Jeffrey's divergence between moving-average models that are real or complex, noise-free or disturbed by additive white noises," *Signal Processing*, vol. 131, pp. 350–363, 2017.

[20] K. K. Sharma, A. Seal, "Modeling uncertain data using monte carlo integration method for clustering," *Expert systems with applications*, vol. 137, pp. 100–116, 2019.

[21] A. Seal, A. Karlekar, O. Krejcar, C. Gonzalo-Martin, "Fuzzy c- means clustering using jeffreys-divergence based similarity measure," *Applied Soft Computing*, vol. 88, p. 106016, 2020.

[22] F. Nielsen, R. Nock, "Total jensen divergences: Definition, properties and clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 2016–2020.

[23] F. Nielsen, R. Nock, S. I. Amari, "On clustering histograms with k-means by using mixed -divergences," *Entropy*, vol. 16, 2014.

[24] R. Nock, F. Nielsen, S.-I. Amari, "On conformal divergences and their population minimizers," *IEEE Transactions on Information Theory*, vol. 62, 2016.

[25] M. D. Gupta, S. Srinivasa, J. Madhukara, M. Antony, "Kl divergence based agglomerative clustering for automated vitiligo grading," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2700–2709.

[26] A. Notsu, O. Komori, S. Eguchi, "Spontaneous clustering via minimum gamma-divergence," *Neural Computation*, vol. 26, 2014.

[27] K. K. Sharma, A. Seal, "Clustering analysis using an adaptive fused distance," *Engineering Applications of Artificial Intelli- gence*, vol. 96, p. 103928, 2020.

[28] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, 2010.

[29] L. Legrand, E. Grivel, "Jeffrey's divergence between moving-average models that are real or complex, noise-free or disturbed by additive white noises," *Signal Processing*, vol. 131, 2017.

[30] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, 2011.

[31] D. Dheeru, E. Karra Taniskidou, "Uci machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[32] U. Maulik, S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, 2002.

[33] N. X. Vinh, J. Epps, J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, no. Oct, 2010.

[34] L. Hubert, P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, 1985.

[35] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[36] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probabil- ity*, vol. 1, 1967, pp. 281–297, Oakland, CA, USA.

[37] S. Chakraborty, S. Das, "k- means clustering with a new divergence-based distance metric: Convergence and performance analysis," *Pattern Recognition Letters*, vol. 100, pp. 67–73, 2017.

[38] L. Hubert, P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

Ayan Seal

Ayan Seal received the PhD degree in engineering from Jadavpur University, West Bengal, India, in 2014. He is currently an Assistant Professor with the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Madhya Pradesh, 482005, India. He is also associated with Center for Basic and Applied Research, Faculty of Informatics and Management, University, of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, 50003, Czech Republic, He has visited the Universidad Politecnica de Madrid, Spain as a visiting research scholar. He is the recipient of several awards. Recently, he has received Sir Visvesvaraya Young Faculty Research Fellowship from Media Lab Asia, Ministry of Electronics and Information Technology, Government of India. He has authored or co-authored of several journals, conferences and book chapters in the area of biometric and medical image processing. His current research interests include image processing and pattern recognition.

Aditya Karlekar

Aditya Karlekar received the B.Tech degree in engineering from Hitkarini College of Engineering and Technology, Jabalpur, 482005. He is currently an intern with the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Madhya Pradesh, India. His current research interests include image processing and pattern recognition.

Ondrej Krejcar

Ondrej Krejcar is a full professor in systems engineering and informatics at the University of Hradec Kralove, Czech Republic. In 2008 he received his Ph.D. title in technical cybernetics at Technical University of Ostrava, Czech Republic. He is currently a vice-rector for science and creative activities of the University of Hradec Kralove from June 2020. At present, he is also a director of the Center for Basic and Applied Research at the University of Hradec Kralove. In years 2016-2020 he was vice-dean for science and research at Faculty of Informatics and Management, UHK. His h-index is 19, with more than 1250 citations received in the Web of Science. In 2018, he was the 14th top peer reviewer in Multidisciplinary in the World according to Publons and a Top Reviewer in the Global Peer Review Awards 2019 by Publons. Currently, he is on the editorial board of the MDPI Sensors IF journal (Q1/Q2 at JCR), and several other ESCI indexed journals. He is a Vice-leader and Management Committee member at WG4 at project COST CA17136, since 2018. He has also been a Management Committee member substitute at project COST CA16226 since 2017. Since 2019, he has been Chairman of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic as a regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic (2019-2024). Since 2020, he has been Chairman of the Panel 1 (Computer, Physical and Chemical Sciences) of the ZETA Program, Technological Agency of the Czech Republic. Since 2014 until 2019, he has been Deputy Chairman of the Panel 7 (Processing Industry, Robotics, and Electrical Engineering) of the Epsilon Program, Technological Agency of the Czech Republic. At the University of Hradec Kralove, he is a guarantee of the doctoral study program in Applied Informatics, where he is focusing on lecturing on Smart Approaches to the Development of Information Systems and Applications in Ubiquitous Computing Environments. His research interests include Control Systems, Smart Sensors, Ubiquitous Computing, Manufacturing, Wireless Technology, Portable Devices, biomedicine, image segmentation and recognition, biometrics, technical cybernetics, and ubiquitous computing. His second area of interest is in Biomedicine (image analysis), as well as Biotelemetric System Architecture (portable device architecture, wireless biosensors), development of applications for mobile devices with use of remote or embedded biomedical sensors.

Enrique Herrera-Viedma

Enrique Herrera-Viedma received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 1993 and 1996, respectively. He is a Professor of computer science and the Vice-President for Research and Knowledge Transfer with University of Granada, Granada, Spain. His h-index is 81 with more than 23 000 citations received in Web of Science and 97 in Google Scholar with more than 37000 cites received. He has been identified as one of the world’s most influential researchers by the Shanghai Center and Thomson Reuters/Clarivate Analytics in both computer science and engineering in the years 2014, 2015, 2016, 2017, 2018, 2019 and 2020. His current research interests include group decision making, consensus models, linguistic modeling, aggregation of information, information retrieval, bibliometric, digital libraries, web quality evaluation, recommender systems, and social media. Dr. Herrera-Viedma is Vice President for Publications in System Man & Cybernetic Society and an Associate Editor in several journals such as IEEE Transactions on Fuzzy Systems, IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE Transactions on Intelligent Transport System, Information Sciences, Applied Soft Computing, Soft Computing, Fuzzy Optimization and Decision Making, and Knowledge-Based Systems.

Optimized DWT Based Digital Image Watermarking and Extraction Using RNN-LSTM

R. Radha Kumari[1]*, V. Vijaya Kumar[2], K. Rama Naidu[3]

[1] Research Scholar, JNT University, Ananthpuramu (India)
[2] Dean, Department of CSE & IT and Director CACR, Anurag Group of Institutions, Hyderabad (India)
[3] Professor, Department of ECE, Jawaharlal Nehru Technological University, Ananthpuramu (India)

Received 27 November 2020 | Accepted 21 August 2021 | Published 27 October 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

The rapid growth of Internet and the fast emergence of multi-media applications over the past decades have led to new problems such as illegal copying, digital plagiarism, distribution and use of copyrighted digital data. Watermarking digital data for copyright protection is a current need of the community. For embedding watermarks, robust algorithms in die media will resolve copyright infringements. Therefore, to enhance the robustness, optimization techniques and deep neural network concepts are utilized. In this paper, the optimized Discrete Wavelet Transform (DWT) is utilized for embedding the watermark. The optimization algorithm is a combination of Simulated Annealing (SA) and Tunicate Swarm Algorithm (TSA). After performing the embedding process, the extraction is processed by deep neural network concept of Recurrent Neural Network based Long Short-Term Memory (RNN-LSTM). From the extraction process, the original image is obtained by this RNN-LSTM method. The experimental set up is carried out in the MATLAB platform. The performance metrics of PSNR, NC and SSIM are determined and compared with existing optimization and machine learning approaches. The results are achieved under various attacks to show the robustness of the proposed work.

Keywords

Discrete Wavelet Transform, Recurrent Neural Network-Long Short-Term Memory, Simulated Annealing, Tunicate Swarm Algorithm, Watermarking.

DOI: 10.9781/ijimai.2021.10.006

I. Introduction

With the rapid growth of digital data use and internet prevalent use, there are frequent acts of intellectual property infringement rights like copying, digital content theft and illegal use [1]. Digital images must be protected because they have high-value added contents for intellectual property rights. The digital watermarking technique is a recently developed technique to protect the digital images. In watermarking, the holder's information embeds the watermark into the content, which is then distributed or stored [2-3]. This topology claims the ownership by extracting embedded watermark information when needed. Numerous schemes have been researched based on the application side, technologies etc. Still, in recent times, various methods have been proposed to extract or modify the watermark algorithm along with the embedding process, to make it a watermark embedding algorithm [4]. In many areas, multimedia copyright protection has been protected by the watermarking techniques for security. Copyright protection is used for different applications like authentication, broadcast monitoring, cryptography and captioning.

To keep the audio, image and video info, the general information embedding technique of watermarking is to be used. It integrates the vital information into methods by invisibly altering the data [5]. Hence, robustness and invisibility are the two significant metrics for estimating the effectiveness of watermarking methods. The watermarking techniques are also categorized into robust watermarking, fragile and semi fragile groups [6]. For data protection of image, the robust watermarking is important as it does not considerably decrease the watermarked image's visual quality and can tolerate numerous attacks. Hence, this is mostly utilized for owner verification and copyright protection. Fragile watermarking is used only to approve the entirety of the image without verifying the actual copyright [7]. For invisibility, the usual method embeds the watermark in DWT, Discrete Cosine Transform (DCT) and Discrete Fourier Transform (DFT) domain.

By comparing the spatial domain and the DCT based watermarking techniques, the DCT based watermarking approaches are robust. The algorithms of the DCT are robust in contrast to modest image processing tasks such as contrast adjustment, low-pass filtration, blurring and brightness, etc [8]. But DCT techniques are hard to implement and computationally expensive. They are also vulnerable to geometric attacks such as cropping, scaling, rotation, etc. It can be categorized into 8×8 block based DCT and global DCT watermarking [9]. Also it was affected by visual part of the image in the low-frequency sub band. Another fact is that the image's high-frequency components are normally detached by noise attacks and compression. Therefore the DCT watermark is embedded by changing the coefficients of the middle frequency sub bands [10]. DFT provides results in the amount of frequency content based on the phase and the magnitude. This is essential for the analysis of images and signals. DWT is, at present, in a variety of signal processing applications like noise mitigation in audio, simulation of wireless antenna distribution and audio and video compression [11]. On wavelets, basic functions are used to represent the signal. DWT is best suited for identifying the regions in the host

* Corresponding author.

E-mail address: radharavada999@gmail.com

image, where the watermark image can be embedded. Wavelets accumulate their energy over time, and they are very suitable for the analysis of transient and time varying signals.

The process of digital watermarking has three steps. In the first stage, the watermark is inserted in embedding. The second step is the attack step, where the degeneration of the watermarked image may or may not occur with a third-party attack. The last step is to find the original image, where the original image is retrieved from the watermarked image. Digital watermarking has been a moderately innovative field during the past two decades. All the digital info can be embedded in the data and then extracted [12]. The information of watermarking can be numbers, handwritten signatures, logos, texts, and can have many uses. It is important to note that one of the basic requirements for a digital watermark is to maintain a quality that does not distort the original data when a watermark is embedded in it. Besides, there are other necessities for specific applications [13].

In recent years, with the development of deep learning [14], convolution Neural Network (CNN) has made advances in the tasks of computer visions like semantic segmentation detection and classification [15]. In the field of super-resolution, the main feature of CNN-based methods is the ability to directly match the complex mapping amongst the low resolution and high resolution images, which enable a better retrieval of the lost high frequency information, and therefore its performance is beyond many classic approaches. Embedding performance is the probability of detection immediately after embedding [16]. While 100% performance is always desirable, it often comes at a very high cost in terms of other features. Depending on the application, one may be willing to sacrifice some performance for a better performance depending on some other characteristics [17].

The main contribution of the work is as follows.

- DWT is applied to protect the images from any attack in terms of embedding the watermark image into the original image.
- A hybrid optimization algorithm SA-TSA is presented to optimize the DWT coefficients on embedding strategy.
- The watermark image extraction process is performed based on a novel deep neural network concept called Recurrent Neural Network based Long Short-Term Memory (RNN-LSTM).
- The proposed method is evaluated and compared with existing methods like Artificial Neural Network (ANN), Deep Neural Network (DNN) and other optimization algorithms.

The rest of this paper is organized as follows: the recent research methods are given in section II. The proposed model description is mentioned in section III. The proposed work algorithm and performance analysis are given in sections IV and V respectively. To end with, the proposed work is concluded in the conclusion section.

II. Related Works

Some of the recent related works are discussed below.

Tanya and Azi [18] have suggested a blind image watermarking approach. The efficacy of the proposed scheme was enhanced by the hybrid Singular Value Decomposition (SVD) and DWT. DWT based 2-level SVD constraints were eliminated by this suggested approach. In watermarking, the watermark size depends on the cover image size. So, the suggested scheme is enhanced to be independent to form cover image sizes and watermark image sizes along with enhancing the imperceptibility and the robustness. For embedding the watermark, image blocking has to discover the optimal size of block and adjust the capability with the host image size. Before watermark embedding, additional pre-processing is done to achieve the goals. Based on the cover image size, the watermark image was discarded. The host and the watermark images were divided into 16×16 blocks. Furthermore, security was ensured by proposing a 2-level authentication method on behalf of extraction to discover the false negative and the false positive issues. The recommended method was tested and practiced for clinical & non-clinical images revealing 10 kinds of geometric attacks.

In this digital period, illegal redistribution and the security of multimedia have become an important issue. Hence, digital watermarking has been presented to avoid the illegal activities and to ensure the authentication and security. Alotaibi [19] has formulated the video watermarking framework which includes video frame prediction by optimal, embedding and extraction processes. By utilizing the Jaya plus firefly algorithm, the optimal frames were selected based on the maximum Peak-signal to-noise-ratio (PSNR). Further frames are allotted with one or zero label, where zero indicates the decreased PSNR and one indicates the better PSNR. As a result, a data library was created since the results obtained, where every frame video was found by its gray-level co-occurrence matrix labels and features. The deep belief network was utilized to optimally select the frame by trained data which enhanced the prediction accuracy and end with watermark embedding and extraction.

In digital watermarking, the quality of digital content should not be compromised and it should not be visible to the human eye. Garg and Kishore [20] have offered a secure watermarking technique for color images. The Particle Swarm Optimization (PSO) algorithm was utilized in the watermarking process. Color watermark was utilized for embedding and embedded in R, G, and B planes of the color input image with entropy encryption and 2D-DCT. By using PSO, the embedding strength was optimized which offers a balance amongst the complexity and the strength. The suggested model was tested by various attacks on watermark image and the performances measures were NC and PSNR. With the intention of contributing to the medical image security, Kahlessenane et al. [21] have presented a robust watermarking approach which consents electronic patient record to be integrated into a computerized tomography scan. Before the integration process, the DWT was applied and the spatial rearrangement of LL subband coefficients was completed by the method of zig zag scanning. The gained coefficients were pooled to integrate the bits of the watermark. The integrity of watermark can be easily verified as it is integrated into the hash image of electronic-patient record. The experimental outcomes gave better PSNR results and good robustness for various attacks. Liu et al. [22] have proposed an image watermarking scheme by using DWT, SVD and Hessenberg decomposition (HD).

First, during the embedding progression, the input image was decomposed into multiple subbands by DWT, and the coefficients were used as HD input. At the same time, the watermark was functioned in the SVD. The watermark was lastly embedded by the scaling factor in the host image. The Fruit Fly optimization algorithm was dedicated for detecting the scaling factor via the objective function estimation process. Furthermore, the suggested technique was associated with the existing methods with various attacks like sharpening, JPEG compression, noise, filter and JPEG2000 compression and the results showed the robustness of the scheme. The transfer of patient record over the network requires a mechanism to ensure the privacy and security of the tele-health services. Ashima and Singh [23] have provided an advanced technique capable of protecting the data of the patients through embedding multiple watermarks on the medical image with the DWT-SVD domain. Before embedding, the hamming code was utilized in the text watermark to decrease the noise distortion on behalf of sensitive data. After the completion of embedding process, the watermarked image was extracted and compressed later. In the experiment, three compressions and two encryption schemes were tested. The outcomes demonstrated that the proposed method provided greater robustness compared to the other methods.

Digital watermarking applications are evidencing the digital content authenticity. For this concern, the watermarking methods were combined to the concepts of artificial networks and histogram. The histogram shape perception was put into practice, with the aim of maintaining the relevance and resistance of the watermark information within the host image. The optimization of the extraction process was proposed in [24]. The suggested artificial neural network was utilized to solve the strengthening resistance problems. The extraction progression was skilled by the auto encoder neural network and the back propagation neural network. The recommended method was checked under numerous attacks which gave better performances. The Firefly Algorithm (FA) is a newly created nature-inspired algorithm that is inspired by the luminous behavior of fireflies so that one firefly tends to attract other fireflies with greater brightness. The benefits of FA have automatic regrouping and local attractions. Hence, Guo et al. [25] has suggested FA based watermarking method with the help of DWT-QR transform. The objective function of the process has a bit error rate and Structural Similarity Index (SSIM). Experimental validation showed the invisibility property and the robustness.

The process of medical images denoising was considered to be a long established setback in the field of image processing. Rajeev et al. [26] have proposed an excellent system for denoising to eliminate salt & pepper and white noises by relating LSTM or RNN and batch normalization technique. The input image of the suggested method was lung CT image. The PSO algorithm was utilized to calculate the batch size in effective way. RNN was utilized to denoise the image. LSTM-based Batch Normalization was presented to minimize the neural network's internal covariate-shift. The assessment matrices were MSE, Signal Noise Ratio or PSNR. The general idea of watermarking work was manipulating the color image underneath various attacks analyzed via the neural network tactic [27]. This was felt over the region of the transformation, particularly an emphasis on contourlet transform to discourse the suggested method, until the bands of the appropriate coefficients were precisely selected. On the color image, the logo information was embedded in the edges, whereas the Zenzo Edge detector was felt to handle the methodology. Actually, the margin of second sub-band was obtained, and then the capacity of the above-referenced edge was computed.

The approaches to embedding and extracting during the learning process of the aforementioned neural network via the training data set are deliberated for continuous research with different scenarios in the current research. The proposed method's effectiveness is verified by the various scenarios analysis obviously.

III. Proposed Watermarking Scheme

The proposed digital image watermarking process is performed based on the process of embedding and extraction. In the process of embedding, the input image is applied to the DWT. In order to improve the robustness of the proposed technique, the DWT wavelet coefficients are optimized with the help of hybrid optimization technique. The SA algorithm is combined with TSA for better optimization. In embedding, the watermark image is converted into bits of watermarks for watermark embedding and to obtain the watermarked image. After performing the embedding process, the extraction is processed by deep neural network concept of Recurrent Neural Network based Long Short-Term memory. Finally, the watermark image and the original image are retrieved by this proposed RNN-LSTM approach. Fig. 1 illustrates the proposed scheme block diagram.

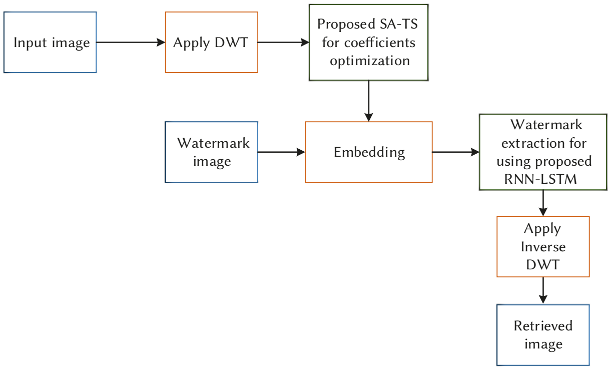

Input image
Apply DWT
Proposed SA-TS for coefficients optimization
Watermark image
Embedding
Watermark extraction for using proposed RNN-LSTM
Apply Inverse DWT
Retrieved image

Fig. 1. Proposed Block Diagram.

A. Watermark Embedding Process

The proposed scheme has utilized the frequency domain; it is in the insertion of the tag, not directly into the image, however into its transformation. Therefore, before the integration, the DWT is applied into the image. The embedded watermarked image is obtained at the final stage of integration. The DWT process is explained below.

1. Discrete Wavelet Transform (DWT)

Many of the watermarking topologies are done by wavelets for its robustness and efficiency contrary to attacks. The image has a low frequency variation with fine details amongst high frequency variations. The low-frequency portion of DWT has high energy and it is more implored than the high frequency portion and separates the low-frequency components from the high-frequency components. In DWT, an image is decomposed into four types of subbands as LL approximation subband, LH, HH and HL detail subbands.

Here, L indicates the low-pass filter and H indicates the high-pass filter which is applied in individually as rows and columns.The choice of subband utilized on behalf of the watermark integration is influenced by the category of images used and its applications. If decomposition is completed, the input image is decomposed into four sub-bands. In contrast, the energy of HL, LH, and HH bands are low. For watermark embedding, the LL sub-band gives better results than the other. Different types of wavelet families are there which depends upon the mother wavelet choice. In our work, the DWT Haar filter is utilized by means of various decomposition levels so as to select the optimum DWT. The DWT is obtained by using the filter of Haar. The Haar filter is utilized to process the signals, and the coefficients of each sub-band is calculated by the equations (1), (2), (3) & (4),

$$LL(a,b) = \frac{p(a,b)+p(a,b+1)+p(a+1,b)+p(a+1,b+1)}{2} \tag{1}$$

$$LH(a,b) = \frac{p(a,b)+p(x,y+1)+p(a+1,y)-p(a+1,b+1)}{2} \tag{2}$$

$$HL(a,b) = \frac{p(a,b)-p(a,b+1)+p(a+1,b)-p(a+1,b+1)}{2} \tag{3}$$

$$HH(a,b) = \frac{p(a,b)-p(a,b+1)-p(a+1,y)+p(a+1,b+1)}{2} \tag{4}$$

Where, image sample pair is denoted as *a, b*. The transform of Haar wavelet decomposes the signal into two signals of half its length, one signal is running on average and other one is running on difference. Equation (5) describes the Haar wavelet mother function $\psi(t)$:

$$\psi(t) = \begin{cases} 1, 0 \le t < 1/2 \\ -1, 1/2 \le t < 1 \\ 0, otherwise \end{cases} \tag{5}$$

The scaling function is described as in equation (6)

$$\phi(t) = \begin{cases} 1, 0 \le t < 1 \\ 0, otherwise \end{cases} \tag{6}$$

Where, the scaling function and the Haar wavelet mother function are denoted as $\phi(t)$ and $\psi(t)$ respectively. The DWT is suitable for the watermark embedding process but the DWT operation enhancement is necessary. Hence, the DWT operation can be improved by manipulating the wavelet coefficients. In our work, the optimal wavelet coefficients are selected by using the proposed Hybrid optimization algorithm as SA-TSA. The following section explains the optimization process.

IV. Optimization Process

In this article, we present the hybrid optimization approach for DWT wavelet coefficient optimization. The SA [28] and the TSA [29] are combined to form the new hybrid optimization algorithm called as SA-TSA. The TS algorithm is based on the behavior of tunicates for discovering the source of food and these activities are based on swarm intelligence and jet propulsion. Before starting the iteration process, the tunicates populations must be initialized. In search space, the TSA populations are randomly initialized. Here, we are using SA algorithm for initialization purpose, which is to enhance the speed of the convergence and the solution accuracy. SA must be applied for every individual of the initial population. In TS algorithm, the population of tunicates consists of $\vec{P_p}$. The objective function of this hybrid optimization is maximizing the PSNR of DWT.

The SA algorithm parameters such as the number of transitions and the temperature are represented as N_n and T_n respectively and n is denoted as iteration. The presented random number lies in between 0 and 1. Hence, the optimization variable of the lower boundary and the upper boundary are indicated as L_b and H_b respectively. It is very important to highlight that each x_i is a vector that is composed of variables that are being optimized. After the initialization, the tunicates swarm behavior and jet propulsion is applied to the whole population.

Basically, the swarm behavior and jet propulsion mechanism of tunicates occurs in four phases as,

i) Avoiding the conflicts among search agents

ii) Association towards the direction of best neighbor

iii) Converge to the finest search agent

iv) Swarm behavior

Step 1: After initialization, each search agent's fitness is calculated.

Step 2: The conflicts between the other tunicates can be avoided by employing the vector $\vec{K}$ as given in equation (7),

$$\vec{K} = \frac{\vec{G}}{\vec{S}} \tag{7}$$

Where, $\vec{G} = k_2 + k_3 - \vec{F}$ and $\vec{F} = 2.k_1$

Here, gravity force and water flow advection is represented as $\vec{G}$ and $\vec{F}$ respectively. The random variables of k_1, k_2, and k_3 lies in the range of zero to one and the social forces among the search agents are denoted as $\vec{S}$, which is given as in equation (8),

$$\vec{S} = T_{min} + k_1.T_{max} - T_{min} \tag{8}$$

Where, social interaction is to make the initial and the subordinate speeds which are denoted as T_{min} and T_{max} respectively. The values of T_{min} and T_{max} are one and four respectively.

Step 3: Distance between the food source and the search agent ($\vec{D}$) is calculated by equation (9),

$$\vec{D} = |F_S - k.\vec{P_p}(x)| \tag{9}$$

Where, x represents the present iteration and the food source location is $\vec{F_s}$ which is the optimal solution. The tunicate location is denoted as $\vec{P_p}(x)$

Step 4: Update the search agent position by equation (10),

$$\vec{P_P}(x') = \begin{cases} \vec{F_S} + \vec{K}.\vec{D}, if k \geq 0.5 \\ \vec{F_S} - \vec{K}\,.\,\vec{D}\,, if k < 0.5 \end{cases} \tag{10}$$

Where, updated tunicate position is denoted as $\vec{P_p}(x')$.

Step 5: Based on the tunicate swarm behavior, the optimal two solutions are stored and updated by the other search agent positions according to the greatest search agent position. The swarm behavior can be expressed as in equation (11),

$$\vec{P_p}(x+1) = \frac{\vec{P_p}(x)+\vec{P_p}(x+1)}{2+k_1} \tag{11}$$

Step 6: Updated search agent is adjusted beyond the boundaries in a given search location.

Step 7: The fitness of the updated search agent value is computed. If there is a better solution than the previous optimal solution, then update P_p.

Step 8: If the stopping criterion is reached, the algorithm will be stopped. Or else, repeat the steps 5–8.

Step 9: Return to the greatest optimal solution ever obtained.

Fig. 2 illustrates the proposed algorithm flowchart.

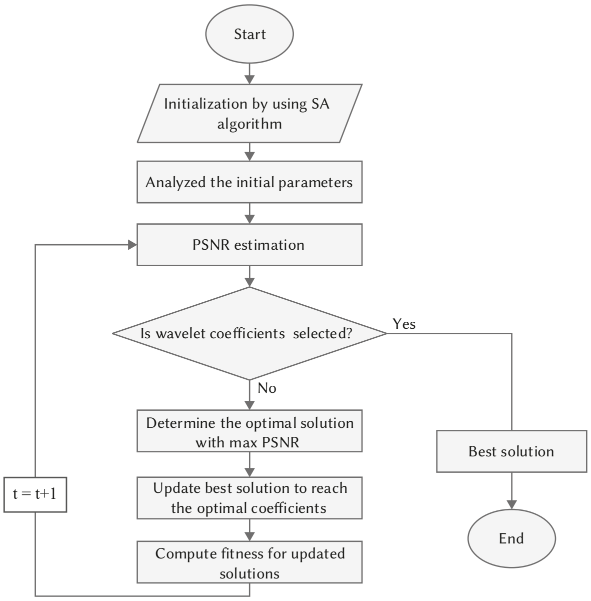

Start
Initialization by using SA algorithm
Analyzed the initial parameters
PSNR estimation
Is wavelet coefficients selected?
Yes
No
Determine the optimal solution with max PSNR
Update best solution to reach the optimal coefficients
Compute fitness for updated solutions
t = t+1
Best solution
End

Fig. 2. Flow chart of the proposed SA-TS algorithm.

Pseudo code for SA algorithm

```
Algorithm 1: Simulated Annealing
For each individual $x_i$
Input data $n = 0$, $T_n = T_0$, $N_n = N_0$
                $x_i = rand \times (H_b - L_b) + L_b$
Repeat
  For $i = 0$ to $N_0$
    Generate solution $x_j$ from the present solution neighborhood $x_i$
      If $f(x_j) < f(x_i)$ then $x_j$ develops the present solution $(x_i = x_j)$
    Else
      $x_j$ develops the present solution with probability
      $e^{\left(\frac{f(x_i)-f(x_j)}{T_n}\right)}$
    $n = n + 1$
calculate $N_n$ and $T_n$
until $T_n \cong 0$
```

The following steps are to be used to embed the digital image watermark embedding:

Step 1: The original image is read at initially.

Step 2: DWT is applied into the input image.

Step 3: To obtain four types of sub bands, the one-level decomposition is performed.

Step 4: The optimized coefficients are obtained by the hybrid SA-TS algorithm.

Step 5: The watermark image is split into many blocks based on the pixel size 64×64.

Step 6: The watermark image is to be embedded and added to the information of sub-band.

Step 7: Obtained the watermarked image by the IDWT performance with an optimized DWT coefficients.

A. Watermark Extraction Using Neural Network

Recurrent Neural Network

An LSTM is designed to work differently than a CNN because an LSTM is usually used to process and make prediction on given sequence of data. During processing, the CNN completely loses all the information about the composition and position of the components and transmits the information further to a neuron which might not be able to extract the image. A Convolutional neural network is significantly slower due to an operation called maxpool. If the CNN has several layers then the training process takes a lot of time if the computer does not consist of a good GPU. Therefore, RNN-LSTM is better over the CNN technique [30].

The neural network of RNN has three layers as input, hidden and status layer. The input vector sequences are $x_1, x_2, ..., x_T$ and the hidden states sequences are $h_1, h_2, ..., h_T$ which are computed by step time t and expressed as in equation (12),

$$h_t = \phi(\omega_h h_{t-1} + \omega_x x_t) \tag{12}$$

Here, the recurrent weight matrix, the hidden weight matrix, and the arbitrary activation function are denoted as ω_h, ω_x and ϕ respectively.

Finally, RNNs stack by inputting h into an additional different RNN, as a result creating deeper structures.

$$h_t^l = \phi(\omega_h h_{t-1}^l + \omega_x h_t^{l-1}) \tag{13}$$

In RNN, the sigmoid function of the activation function is represented by Φ which is a hyperbolic tangent. Networks' training is recognized as particularly difficult because the gradients are explodes and vanish. RNN has a specific memory function; however it cannot overcome the long-term dependency issues due to the complications of gradient explosion, gradient dispersion and RNN training. The LSTM network is a special RNN introduced by Schmidhuber and Hochreiter in 1997 [31]. The LSTM has solved the issues of long-term dependencies and continuously enhanced using the academic community. The LSTM's hidden layer arrangement is the long short-term memory block, thereby the memory block contains cell structure and three thresholds control named as forget gate (f_t), input gate (i_t) and output gate (o_t). By using the input vector (h_{t-1}, x_t), the LSTM memory cell state is calculated by the forget gate.

$$f_t = \sigma(\omega_f . [h_{t-1}, x_t] + b_f) \tag{14}$$

The new information generated in next part requires to be updated. This is considered as two stages; at first, the input gate computes the values throughout the used sigmoid function. Next, the new candidate values c_t is generated by that layer, which is added into the memory cell.

Where, the current time input and the last time output are denoted as x_t and h_{t-1} respectively. The input layer bias and the weight are denoted as ω_f and b_f and activation function is $\sigma(.)$ which is generally known as sigmoid function.

$$i_t = \sigma(\omega_i . [h_{t-1}, x_t] + b_i) \tag{15}$$

$$\tilde{c}_t = tanh(\omega_c . [h_{t-1}, x_t] + b_c \tag{16}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{17}$$

Here, the status update layer bias and the weight function are denoted as b_c and ω_c respectively.

Finally, the network is controlled by the updated state through the output gates in the output layer which is described in equation (18),

$$o_t = \sigma(\omega_o [h_{t-1}, x_t] + b_o) \tag{18}$$

$$h_t = o_t * tanh(c_t) \tag{19}$$

Where, the output layer bias and the weight are denoted as b_o and ω_o respectively.

B. Steps for Watermark Extraction

Step 1: Forward DWT of the watermarked image taken.

Step 2: The image pixel blocks are selected.

Step 3: For the selected blocks, the middle value in each block is applied to RNN-LSTM.

Step 4: Based on the RNN-LSTM output and other some information, the watermark can be extracted from the watermarked image.

Step 5: Subtracted from the embedding reverse process by using the equation (20),

$$extrc\_wtmk = (emb - mat)/(4 * \mathrm{y} + 2) + 0.5\ (4.9) \tag{20}$$

Where, y is denoted as neural network output and mat is the obtained matrix elements.

Step 6: Extract a watermark image and an original image.

Before extracting the watermark, it is important to distinguish among the watermarked and the original image. This should be done to check the watermarked image strength and compared to the original image. Subtraction of the original image from the watermarked image was finished in this connection. The watermark can be extracted by the second hidden layer of the neural network compared to the pixels of the distorted image pixel. Therefore the visible watermark is extracted by the proposed RNN-LSTM network.

V. Results and Discussion

The proposed work is implemented in the platform of MATLAB Intel core3 processor 2018b version. For the digital watermarking technique, the DWT was utilized to embed the watermark in the host image and the embedded watermark is robust against the various types of attacks. From the watermarked image, the original input image is extracted by the utilization of RNN-LSTM concept. For the experimental setup, the various test images are taken. The test images are peppers, Barbara, Lena and person and the image size is 512×512. We consider two watermark images; one is the copyright logo and another one is the cameraman image which has 64×64 pixel size. Fig. 3 and Fig. 4 show the original image and watermark image, respectively.

Performance metrics

To maintain the accuracy, we have used the following parameters as PSNR, Normalized coefficient (NC) and Mean Square Error (MSE). These matrices are evaluated in the proposed work.

a) Peppers b) Barbara c) Lena d) Person

Fig. 3. Original Image.

a) Copyright logo b) Cameraman

Fig. 4. Watermark Image.

Peak signal to noise ratio

The watermarked image quality is measured by the parameter of PSNR. The ratio between the input image and the watermarked image is known as PSNR. Also the PSNR is measured based on the MSE. The PSNR is defined as in equation (21),

$$PSNR = 10\, log_{10}(\frac{255^2}{MSE}) \tag{21}$$

Normalized correlation

It is used to evaluate the distance between the two vectors. It can be defined as in equation (22),

$$NC = \sum_{a=1}^{n}\sum_{b=1}^{n} I_{(a,b)}^{in} XOR I_{(a,b)}^{w} / n * n \tag{22}$$

The input image is denoted as $I_{(a,b)}^{in}$ and the watermarked image is denoted as $I_{(a,b)}^{w}$. For watermark bit generation, the Exclusive-OR (XOR) operation is used in watermark image embedding.

Structural Similarity Index Measure

The frame or image quality prediction method is called as SSIM. The SSIM is computed by the frame or images' various windows size. The SSIM is described as in equation (23),

$$SSIM(X,Y) = \frac{(\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{23}$$

The input images' pixel size is 512 × 512 and the watermark images' pixel size is 64 × 64. The watermark image is converted into binary bits. By using the DWT, the decomposition is performed by the Haar filter. The DWT four level coefficients are optimized by the proposed SA-TSA algorithm. For each subsequent stage of wavelet decomposition, the previous level LL subband is utilized as input. At last, three levels of four sub-bands are attained, each band's pixel is 64 × 64. The host input image is reconstructed by the optimized DWT coefficient which is called as IDWT. Then the 64 × 64 pixel size is allocated into small blocks by 4 × 4 pixel sizes. For this, the wavelet transform incorporates the necessary features to attain the maximum advantages. Then the binary bits watermark image with 64 × 64 pixel is divided into non-overlapping small blocks with 4 × 2 pixel sizes, hence, 16 × 32 blocks are produced. After that, these blocks are inserted into the chosen wavelet coefficient blocks. Then the watermark image is embedded into block by block which reduces the processing time. The proposed image watermarking performance is tabulated in Table I.

For the extraction process, the trained RNN-LSTM network is to be used which is suitable for learning the relation amongst the watermarked image wavelet coefficients and the watermark image corresponding pixel. The decomposed watermarked image and its coefficients are split into 4 × 4 pixel small block dimensions. Then these coefficients' contents are extracted and utilized as the input of the trained RNN-LSTM and to achieve the watermarked data.

A. Robustness of the Proposed Method

Robustness is defined as the ability to attack variation without embracing the initial static formation of a system. In digital image watermarking scheme, robustness means the ability to extract the watermark from a watermarked image under several attacks. Hence, it is significant to identify the robustness of a watermarking system. To detect the proposed method performance, the two types of watermark images (watermark logo and cameraman) have been utilized to compare the robustness of the proposed techniques.

The recommended image watermarking approach has been evaluated for the outcomes and the robustness which are validated by the Tables II, III, IV and V respectively. The robustness of the RNN-LSTM network based Peppers, Barbara, Lena and Person images' various attacked environments are presented. For the comparison, the performance matrices PSNR, NC and SSIM are measured.

SSIM is used to compare the watermarked images qualities after applying the attacks. For binary watermark, the NC is used to compare the robustness level. Table II describes the proposed RNN-LSTM network based method that achieves the highest NC value which is something that can be achieved. Therefore, the proposed method performs outstandingly by means of extracting the whole embedding watermark bits.

Table III describes the various attacks for Barbara image. At no attacks, the watermark image 1 PSNR is 56.005 and the watermark image 2 PSNR is 57.26. Thereby the cameraman watermark image performance has achieved a better PSNR. Table IV describes the various attacks for Lena image. At no attacks, the watermark image 1 PSNR is 56.004 and the watermark image 2 PSNR is 54.56. Thereby the copyright logo watermark image performance has achieved a better PSNR. In the comparison of robustness, the proposed method's performances are achieved with the no attacks, speckle noise, salt & pepper noise, Gaussian noise, sharpening attack, rotation attack, motion blur, average filter, Jpeg compression, histogram equalization, rescaling and wiener filter.

Extracted watermark was affected by applying various attacks. If average filtering is applied, the NC value is close to 1 as 0.996. Also, the image quality is affected, wherever the SSIM average values are nearer to 0.994 at watermark 1. In watermark 2, the NC is 0.9960 and SSIM is 0.9923. In the image watermarking, the high PSNR is achieved at no attacks as 56.004. When adding the speckle noise, Gaussian noise and salt & pepper noise, the value of PSNR is decreased at 47.916, 46.30 and 38.503 respectively. In watermark image 2, the PSNR value is 54.56. Similarly, Table V describes the various attacks for Person image.

B. Performance Evaluation

In the proposed image watermarking scheme, from the watermarked image, the watermark image and the original image are extracted by using the RNN-LSTM network. The neural network approaches are suitable for this watermark extraction performance. After the attacks of cropping, rotation and JPEG the extracted images are partially degraded. Nevertheless, the extracted watermark images remain recognizable because these attacks change the indexed reference values, which may contain the watermark location values of the embedded values. If increases the RNN inputs, the watermark image extraction quality is degraded. This effect is seen because the

TABLE I. Performance of the Proposed Image Watermarking

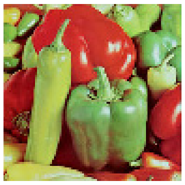
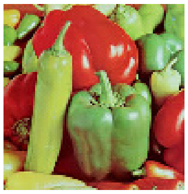
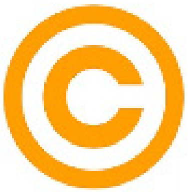
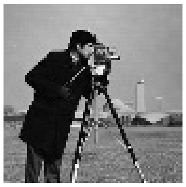
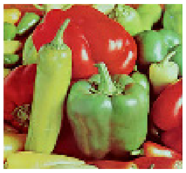
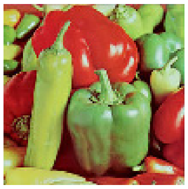
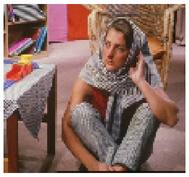
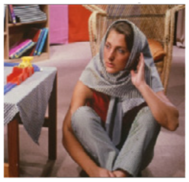
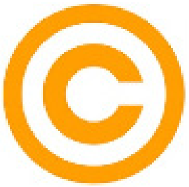
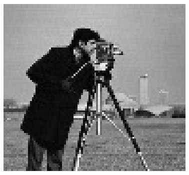
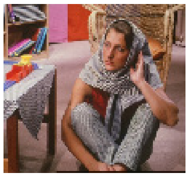
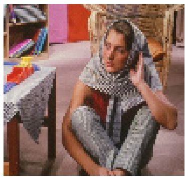
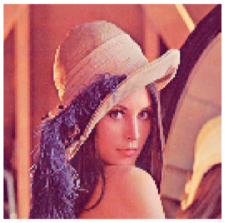
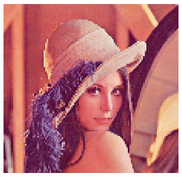
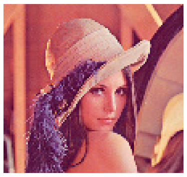
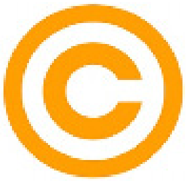
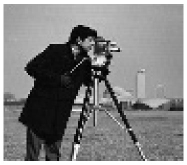
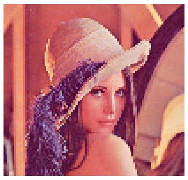
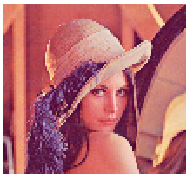
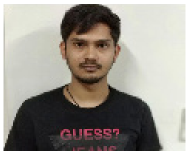
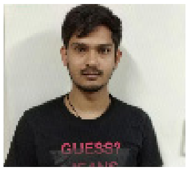
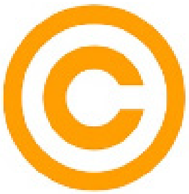
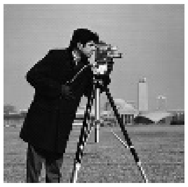
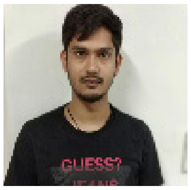
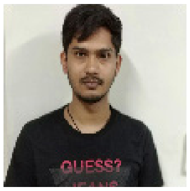
Image	Original images	Watermark image	Watermarked image	Extracted watermark image	Extracted original image
Peppers		 	 	 	 
Barbara		 	 	 	 
Lena		 	 	 	 
Person		 	 	 	 

TABLE II. Various Attacks for Peppers Image

Pepper		Watermark image 1			Watermark image 2		
Image processing attacks	**Value**	**PSNR**	**NC**	**SSIM**	**PSNR**	**NC**	**SSIM**
No attack	-	55.9949	0.9999	0.99718	61.0912	0.9995	0.9889
Speckle noise	0.11	47.5906	0.9984	0.97876	41.2169	0.9999	0.9675
Salt & pepper noise	0.1	45.7953	0.9978	098667	39.9951	0.9961	0.9784
Gaussian noise	0.002	38.6992	0.9987	0.86835	33.1583	0.9963	0.8466
Sharpening attack	0.002	55.3312	0.9968	0.99606	45.234	0.9998	0.9878
Rotation attack	0.002	50.8399	0.9912	0.99244	43.6997	0.9914	0.9844
Motion blur	0.5	54.6194	0.9991	0.99632	47.828	0.9963	0.9950
Average filter	0.8	54.195	0.9989	0.99281	47.5532	0.9978	0.9943
Jpeg compression	50	53.6788	0.9996	0.99508	47.2182	0.9921	0.9936
Histogram equalization	0.5	53.2366	0.9925	0.99247	43.114	0.9939	0.9800
Rescaling	0.25	55.9949	0.9969	0.99718	45.3084	0.9984	0.9888
Wiener filter	0.5	45.9911	0.9992	0.99788	45.3121	0.9971	0.9847

TABLE III. Various Attacks for Barbara Image

Barbara		Watermark image 1			Watermark image 2		
Image processing attacks	**Value**	**PSNR**	**NC**	**SSIM**	**PSNR**	**NC**	**SSIM**
No attack	-	56.0051	0.9989	0.99719	57.26	0.999	0.9822
Speckle noise	0.11	48.0643	0.9985	0.97863	41.3977	0.9942	0.9688
Salt & pepper noise	0.1	46.3968	0.9950	0.98679	40.1401	0.9981	0.9781
Gaussian noise	0.002	38.5094	0.9963	0.86404	32.9651	0.9957	0.8412
Sharpening attack	0.002	55.4092	0.9959	0.95562	45.2324	0.9966	0.9880
Rotation attack	0.002	42.6056	0.9928	0.95562	36.957	0.9909	0.9503
Motion blur	0.5	54.1865	0.9900	0.99567	47.5454	0.9925	0.9942
Average filter	0.8	52.974	0.9904	0.99381	46.6387	0.9932	0.9922
Jpeg compression	50	46.0539	0.9915	0.973	40.4888	0.9947	0.9751
Histogram equalization	0.5	52.565	0.9931	0.99289	42.6127	0.9937	0.9779
Rescaling	0.25	54.2541	0.9907	0.99487	45.2619	0.9901	0.9887
Wiener filter	0.5	50.4452	0.9933	0.9709	46.1529	0.9973	0.9805

TABLE IV. Various Attacks for Lena Image

Lena		Watermark image 1			Watermark image 2		
Image processing attacks	**Value**	**PSNR**	**NC**	**SSIM**	**PSNR**	**NC**	**SSIM**
No attack	-	56.0045	0.9999	0.99719	54.5671	0.9999	0.9887
Speckle noise	0.11	47.9165	0.9967	0.97779	41.2569	0.9951	0.9686
Salt & pepper noise	0.1	46.3007	0.9967	0.98652	40.225	0.9906	0.9776
Gaussian noise	0.002	38.5032	0.9954	0.86394	32.9626	0.9975	0.8412
Sharpening attack	0.002	55.4581	0.9964	0.99625	45.2069	0.9917	0.9878
Rotation attack	0.002	45.277	0.9922	0.96869	39.3983	0.9999	0.958
Motion blur	0.5	54.3272	0.9964	0.99576	47.6103	0.9996	0.9945
Average filter	0.8	53.0465	0.9960	0.994	46.7244	0.9969	0.9923
Jpeg compression	50	50.7845	0.9952	0.98998	44.8585	0.9960	0.9874
Histogram equalization	0.5	50.8462	0.9996	0.99012	41.9573	0.9910	0.9732
Rescaling	0.25	46.5957	0.9989	0.9934	46.6875	0.9909	0.9923
Wiener filter	0.5	50.4436	0.9994	0.9970	45.2626	0.9988	0.9887

TABLE V. Various Attacks for Real-Time Person Image

Person		Watermark image 1			Watermark image 2		
Image processing attacks	**Value**	**PSNR**	**NC**	**SSIM**	**PSNR**	**NC**	**SSIM**
No attack	-	56.0052	0.9999	0.99718	55.6894	0.9999	0.9888
Speckle noise	0.11	45.0682	0.9984	0.96322	39.0039	0.999	0.942
Salt & pepper noise	0.1	45.7215	0.9985	0.98741	39.8337	0.9962	0.9793
Gaussian noise	0.002	38.5538	0.9925	0.86473	33.0254	0.9933	0.8428
Sharpening attack	0.002	55.3819	0.9969	0.99615	45.2583	0.9905	0.988
Rotation attack	0.002	49.7753	0.9981	0.99133	42.912	0.9961	0.9851
Motion blur	0.5	54.8852	0.9952	0.99656	48.0028	0.9925	0.9955
Average filter	0.8	54.6617	0.9947	0.99619	47.8321	0.9967	0.9952
Jpeg compression	50	54.1889	0.9911	0.99542	47.4996	0.9931	0.9949
Histogram equalization	0.5	50.0393	0.9936	0.99059	37.7344	0.9910	0.9335
Rescaling	0.25	46.7290	0.9963	0.9872	45.9545	0.9978	0.9888
Wiener filter	0.5	47.8321	0.9987	0.9953	45.2679	0.9981	0.9887

higher the numbers of RNN-LSTM inputs indicate the fewer neural networks available for watermark extraction, i.e. lower the accuracy and computation time of RNN-LSTM. In watermarking, embedding and extraction computational complexity is necessary. The embedding strategy is easily and quickly performed but extraction is time-consuming process. Excellent watermarks with high NC and PSNR values should be extracted for various aspects of watermarking algorithms, which prioritize image quality over speed. To analyses the effectiveness, the proposed RNN-LSTM neural network is compared with ANN [24] and DNN [32]. ANN, DNN and proposed RNN-LSTM are the deep neural network family techniques. Hence, the proposed method is compared with ANN and DNN. To show the proposed method efficiency, the PSNR, NC and SSIM values are calculated and compared with other deep neural networks.

From the comparison of Fig. 5, 6 and 7, it is clear that the value of proposed method PSNR is high compared to the other methods. Similarly, the value of NC and SSIM is higher than the existing methods. Therefore, RNN-LSTM assists better than the conventional methods in terms of PSNR, NC and SSIM. Table VI shows the measures of PSNR, NC and SSIM. The pepper image gives 55.994 db and 61.09 PSNR for watermark image 1 and watermark image 2. Similarly, for Barbara, Lena and person images, the PSNR value is higher than the other approaches. The NC metrics of the Pepper image value is 1 for watermark image 1 and watermark image 2, the RNN-LSTM NC value is 0.0998 which is higher than the existing methods. From the table VI, Barbara image SSIM value of proposed method is 0.9709 and existing method (ANN) is 0.965 and DNN is 0.965. Likewise, pepper, Lena, and Person images' proposed SSIM value is higher than the existing methods. This performance is verified by the extracted watermark results, which demonstrates that the algorithm can maintain the quality of the watermarked image after embedding. The functional values of the image quality measurement confirm this result. Therefore, the watermarked image created by the proposed method has better imperceptibility than obtained using similar techniques because the PSNR values indicate that the watermarked image and the original image are identical.

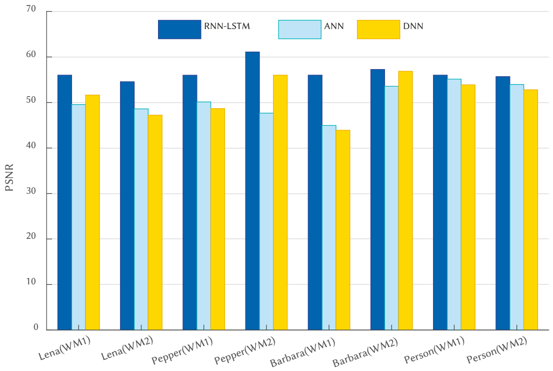


Fig. 5. Comparison of PSNR.

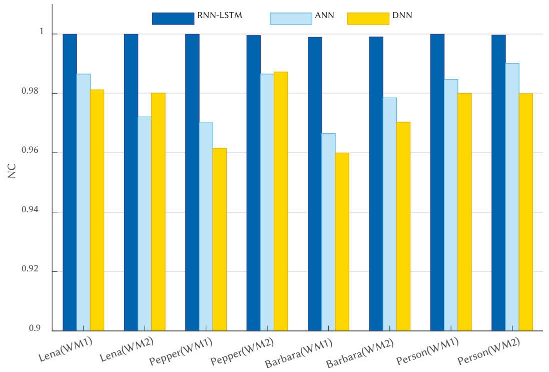


Fig. 6. Comparison of NC.

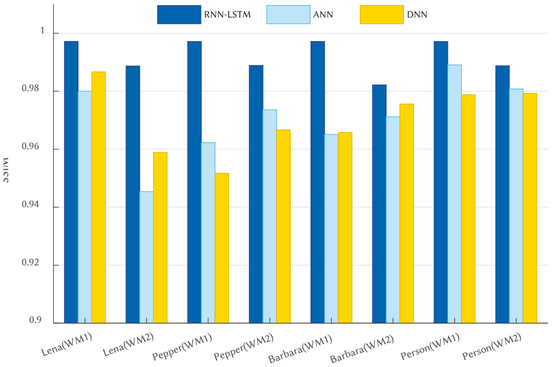


Fig. 7. Comparison of SSIM.

In the proposed work, the DWT wavelet coefficients are optimally selected by the proposed hybrid SA-TS algorithm. For a better optimal solution, we have utilized the hybrid optimization. Based on the maximum PSNR, the optimal coefficients are selected. The robustness of the algorithm is evaluated by the convergence plot of the proposed algorithm. Here the convergence graph is plotted with PSNR and its corresponding iterations. For the evaluation, the proposed SA-TSA is compared with the existing approaches of TSA and PSO [26]. The algorithm parameters are tabulated in Table VII. The three types of input images are taken and also two watermark images are utilized. So, the individual image PSNR convergence performances are displayed in the figures. Fig. 8 shows the Peppers image convergence graph for watermark image 1 and watermark image 2. By comparing the SA-TSA with PSO and TSA, the proposed method is achieving high PSNR (55.99) at watermark image 1. In watermark image 2, the proposed PSNR is 61.09 which are higher than the existing algorithm as TSA (48.01) and PSO (47.23). If the number of iterations increases, the PSNR value is to be maintained constant.

TABLE VI. Comparison Graph of PSNR, NC and SSIM With Existing Techniques

Image		RNN-LSTM			ANN			DNN		
		PSNR	NC	SSIM	PSNR	NC	SSIM	PSNR	NC	SSIM
Peppers	Watermark Img1	55.994	0.999	0.9971	50.122	0.970	0.962	48.659	0.961	0.951
	Watermark Img2	61.091	0.999	0.9889	47.658	0.986	0.973	55.998	0.987	0.966
Barbara	Watermark Img1	56.005	0.998	0.9971	44.958	0.966	0.965	43.900	0.959	0.965
	Watermark Img2	57.26	0.999	0.982	53.552	0.978	0.981	56.854	0.970	0.975
Lena	Watermark Img1	56.004	0.999	0.9919	49.548	0.986	0.98	51.624	0.981	0.986
	Watermark Img2	54.567	0.999	0.9887	48.593	0.972	0.945	47.214	0.980	0.958
Person	Watermark Img1	56.005	0.999	0.9971	55.125	0.984	0.989	53.874	0.980	0.978
	Watermark Img2	55.689	0.999	0.9888	53.958	0.990	0.980	52.789	0.979	0.979

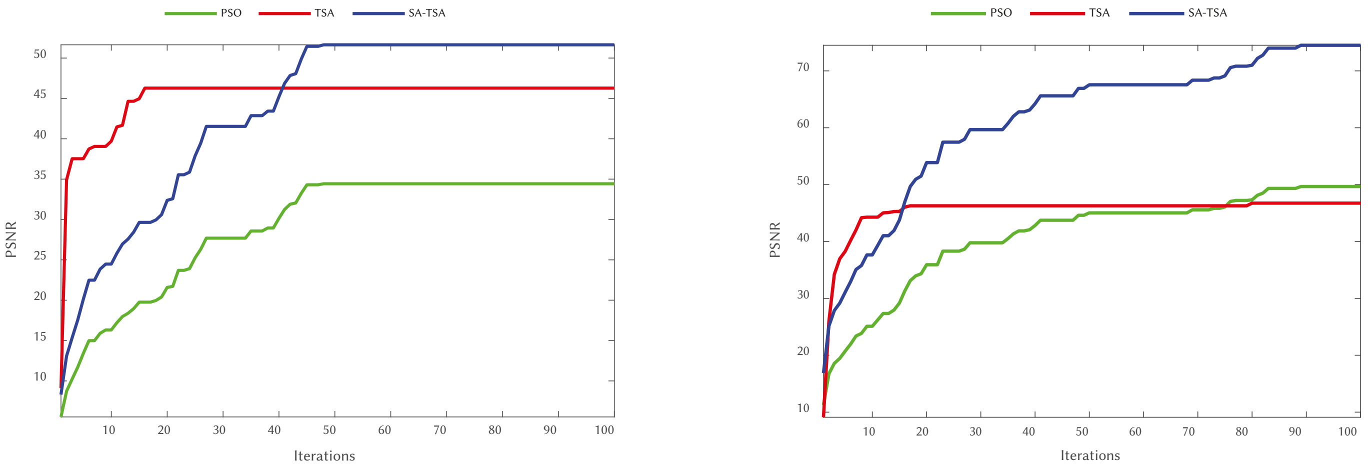
TABLE VII. Algorithm Parameters

Algorithms	Parameters	Values
Proposed SA-TS	Number of iterations	100
	Number of search agents	Size of wavelet coefficient
	Number of population	50
	Dimension	5
	Fitness	Maximum PSNR
TSA	Search agents	20
	I_{min}	1
	I_{max}	4
	Number of iterations	100
PSO	Number of particles	20
	Inertia coefficient	0.75
	Cognitive & social coefficient	1.8 & 2
	Number of iterations	100
RNN-LSTM	Number of hidden units	120
	Filter size	5
	Number of filters	20
	Dropout	0.1
	Mini batch	320
	Optimization	Adam
	Recurrent dropout	0.1
	Loss function	Cross entropy

Fig. 9 shows the Barbara image convergence graph for watermark image 1 & watermark image 2. By comparing the SA-TSA with PSO (40.07) and TSA (48.15), the proposed method is achieving high PSNR (56.005) at watermark image 1. In watermark image 2, the proposed PSNR is 57.26 which are higher than the existing algorithm as TSA (49.46) and PSO (42.56). Fig. 10 shows the Lena image convergence graph for watermark image 1 and watermark image 2. By comparing the SA-TSA with PSO (43.56) and TSA (45.08), the proposed method is achieving high PSNR (56.004) at watermark image 1. In watermark image 2, the proposed PSNR is 54.567 which are higher than the existing algorithm as TSA (50.10) and PSO (45.22). By the comparison, the suggested method gives a better PSNR than the other. Further, the watermark image 2 results are better than the watermark image 1.

Fig. 11 shows the real-time Person image convergence graph for watermark image 1 and watermark image 2. By comparing the SA-TSA with PSO (35.46) and TSA (46.87), the proposed method is achieving high PSNR (56.005) at watermark image 1. In watermark image 2, the proposed PSNR is 55.687 which are higher than the existing algorithm as TSA (47.93) and PSO (46.29). By comparison, the suggested method gives the better PSNR than the other. Further, the watermark image 2 results are better than the watermark image 1.

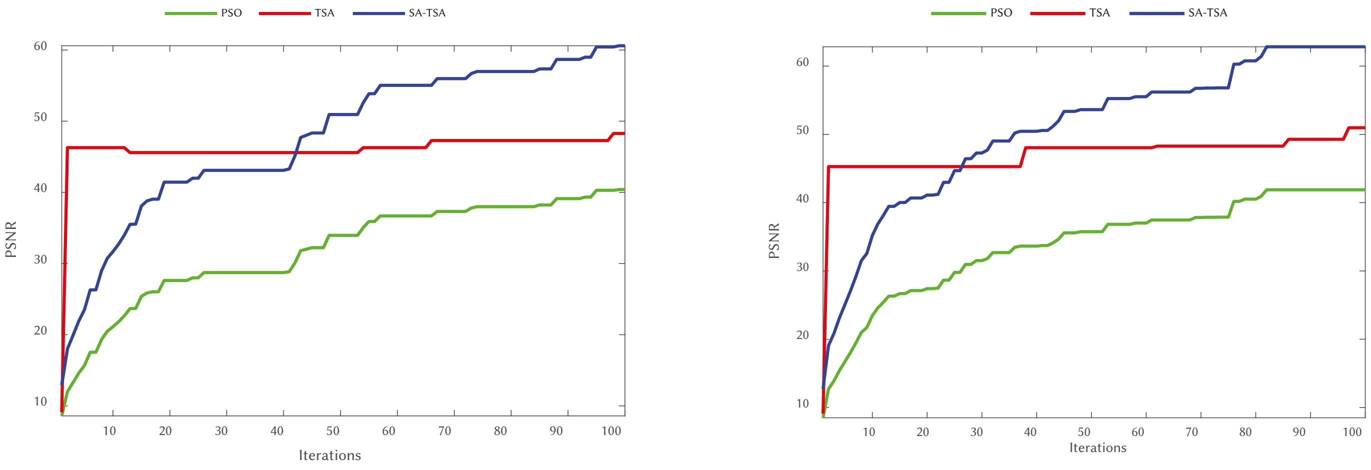
From the outcomes, the recommended digital watermarking approach was performing very well in terms of robustness and



(a)

(b)

Fig. 8. Convergence for Peppers input image: a) watermarked image 1 and b) watermarked image 2.



(a)

(b)

Fig. 9. Convergence for Barbara input image: a) watermarked image 1 and b) watermarked image 2.

performance metrics. For the experimental assessment, the recommended method accomplishes more than 50dB PSNR which evaluates the quality of image watermarking. The robustness of the proposed model accomplished a good result by the SA-TS optimization algorithm and RNN-LSTM. Hence, the proposed method performances are verified in terms of perceptual quality which gives excellent results and achieves enhanced robustness performance.

VI. Conclusion

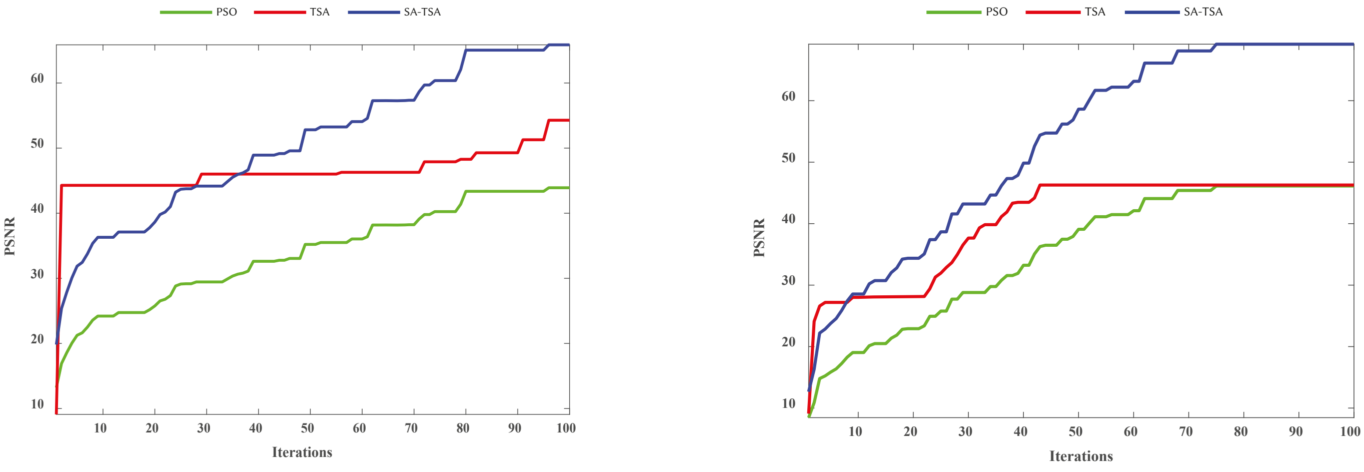
The present research work focused on the development of an imperceptible and a robust non-blind image watermarking. The proposed digital image watermarking method uses DWT, SA-TS algorithm and RNN-LSTM for extraction to achieve the objectives. It is developed to embed a watermark in the host image without noticeable visual artifacts or degradation. The DWT coefficients are optimized by the proposed SA-TS algorithm. The watermark embedding is performed by DWT with SA-TS algorithm. The watermark extraction is done by the machine learning concept of RNN-LSTM. Four input images and two watermark images are provided in the experiment. Under various attacks, the results are achieved to show the robustness of the proposed work. In the comparison of robustness, the proposed method achieves better results with speckle noise, salt & pepper noise, Gaussian noise, sharpening attack, rotation attack, motion blur, average filter, Jpeg compression, histogram equalization, rescaling and wiener filter. The proposed optimization algorithm performance is compared with TSA and PSO algorithm. From the convergence, the maximum PSNR is achieved by the proposed algorithm. Also, the RNN-LSTM results are compared with the measures like PSNR (avg58.542), NC (0.999) and SSIM (avg0.992) with DNN and ANN. From an overall perspective, this technique can bring substantial benefits to the field of digital watermark and additional benefits for copyright protection. In the future work, deep learning network will be used in both embedding and extraction step. Additionally, the image preprocessing approach is used to enhance watermark image quality.

References

[1] I. Hamamoto, M. Kawamura. "Image watermarking technique using embedder and extractor neural networks." *IEICE TRANSACTIONS on Information and Systems*, Vol.102, No. 1, January 2019, pp. 19-30. DOI:https://doi.org/10.1587/transinf.2018MUP0006.

[2] X. Zhou, H. Zhang, and C. Wang. "A robust image watermarking technique based on DWT, APDCBT, and SVD." *Symmetry*, Vol .10, No. 3, March 2018, pp. 77. DOI:https://doi.org/10.3390/sym10030077.

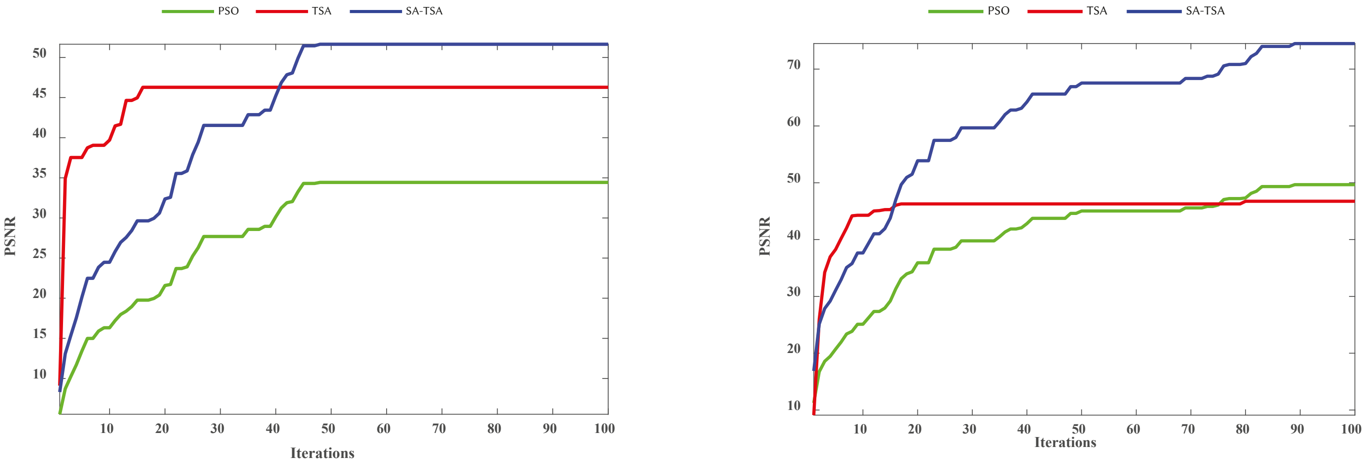
[3] Z. Renjie, X. Zhang, M. Shi, and Z. Tang. "Secure neural network



(a)

(b)

Fig. 10. Convergence for Lena input image: a) watermarked image 1 and b) watermarked image 2.



(a)

(b)

Fig. 11. Convergence for Person input image: a) watermarked image 1 and b) watermarked image 2.

watermarking protocol against forging attack." *EURASIP Journal on Image and Video Processing, September* 2020, pp. 1-12. DOI: https://doi.org/10.1186/s13640-020-00527-1.

[4] S.P. Ambadekar, J. Jain, and J. Khanapuri. "Digital image watermarking through encryption and DWT for copyright protection." *In Recent Trends in Signal and Image Processing*, Vol.727, May 2018, pp. 187-195.

[5] T.K Araghi, A. A Manaf. "An enhanced hybrid image watermarking scheme for security of medical and non-medical images based on DWT and 2-D SVD." *Future Generation Computer Systems* Vol.101, December 2019, pp. 1223-1246. DOI: 10.1016/j.future.2019.07.064.

[6] M. Ali, and C. W. Ahn. "An optimal image watermarking approach through cuckoo search algorithm in wavelet domain." *International Journal of System Assurance Engineering and Management*, Vol.9, No. 3, June 2018, pp. 602-611. DOI: 10.1007/s13198-014-0288-4.

[7] D. Rajani, and P. Rajesh Kumar. "An optimized blind watermarking scheme based on principal component analysis in redundant discrete wavelet domain." *Signal Processing*, Vol. 172, July 2020, pp.107556. DOI: https://doi.org/10.1016/j.sigpro.2020.107556.

[8] X. Kang, F. Zhao, G. Lin, and Y. Chen. "A novel hybrid of DCT and SVD in DWT domain for robust and invisible blind image watermarking with optimal embedding strength." *Multimedia Tools and Applications,* Vol 77, No. 11, July2017, pp. 13197-13224. DOI:https://doi.org/10.1007/s11042-017-4941-1.

[9] P. Kadian, N. Arora, and S. M. Arora. "Performance Evaluation of Robust Watermarking Using DWT-SVD and RDWT-SVD." *In 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN) IEEE,* May 2019, pp. 987-991.

[10] A.K. Abdulrahman, and S. Ozturk. "A novel hybrid DCT and DWT based robust watermarking algorithm for color images." *Multimedia Tools and Applications,* Vol. 78, no. 12, January 2019, pp. 17027-17049, DOI:https://doi.org/10.1007/s11042-018-7085-z.

[11] F.N. Thakkar, and V. K. Srivastava. "Performance comparison of recent optimization algorithm Jaya with particle swarm optimization for digital image watermarking in complex wavelet domain." *Multidimensional Systems and Signal Processing,* Vol. 30, no. 4, 2019, pp. 1769-1791.

[12] M. Ahmadi, A. Norouzi, N. Karimi, S. Samavi, and A. Emami. "ReDMark: Framework for residual diffusion watermarking based on deep networks." *Expert Systems with Applications, Vol.146*, May 2020 pp.113157.

[13] F. López, L. de la Fuente Valentín, and I. S. M. de Mendivil. "Detecting image brush editing using the discarded coefficients and intentions." *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 5, no. 5, 2019, pp.15-21.

[14] N. Saleem, and M. I. Khattak. "Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments." *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 6, no. 1, 2020, pp. 84-90.

[15] J.E. Lee, Y.H. Seo, and D.W. Kim. "Convolutional Neural Network-Based Digital Image Watermarking Adaptive to the Resolution of Image and Watermark." *Applied Sciences, Vol.10*, no. 19, September 2020, pp. 6854. https://doi.org/10.3390/app10196854.

[16] Y.Nagai, Y.Uchida, S.Sakazawa, and S. Satoh. "Digital watermarking for deep neural networks." *International Journal of Multimedia Information Retrieval,* Vol 7, no. 1, February 2018, pp. 3-16. DOI:https://doi.org/10.1007/s13735-018-0147-1.

[17] F. Wang, X. Liu, G. Deng, X. Yu, H. Li, and Q.Han. "Remaining life prediction method for rolling bearing based on the long short-term memory network." *Neural Processing Letters,* Vol. 50, no. 3, March 2019, pp2437-2454. DOI: 10.1007/s11063-019-10016-w.

[18] T.K. Araghi, and A. A. Manaf. "An enhanced hybrid image watermarking scheme for security of medical and non-medical images based on DWT and 2-D SVD." *Future Generation Computer Systems* Vol. 101, December 2019, pp. 1223-1246.DOI: 10.1016/j.future.2019.07.064.

[19] S.S. Alotaibi, "Optimization insisted watermarking model: hybrid firefly and Jaya algorithm for video copyright protection." *Soft Computing*, Vol.24, March 2020, pp. 14809-14823. DOI:https://doi.org/10.1007/s00500-020-04833-8.

[20] P. Garg, and R.R.Kishore. "Optimized color image watermarking through watermark strength optimization using particle swarm optimization technique." *Journal of Information and Optimization Sciences* Vol.41, No.6, 2020, pp1-14.

[21] F. Kahlessenane, A. Khaldi, R. Kafi, and S.Euschi. "A DWT based watermarking approach for medical image protection." *Journal of Ambient Intelligence and Humanized Computing*, August 2020, pp1-8.

[22] J. Liu, Huang, Y. Luo, L.Cao, S. Yang, D.Wei, and R. Zhou. "An optimized image watermarking method based on HD and SVD in DWT domain." *IEEE Access, Vol.* 7, May 2019, pp. 80849-80860. **DOI:** 10.1109/ACCESS.2019.2915596.

[23] A. Ashima, and A.K. Singh. "An improved DWT-SVD domain watermarking for medical information security." *Computer Communications* Vol. 152, February 2020, pp.72-80. DOI: https://doi.org/10.1016/j.comcom.2020.01.038.

[24] R.R. Sunesh Kishore, and A. Saini. "Optimized image watermarking with artificial neural networks and histogram shape." *Journal of Information and Optimization Sciences*, Vol .44, No.7, September 2020, pp. 1597-1613.

[25] Y. Guo, B-Z. Li, and N.Goel."Optimised blind image watermarking method based on firefly algorithm in DWT-QR transform domain." *IET Image processing*, Vol.11, No. 6, June 2017, pp.406-415. **DOI:** 10.1049/iet-ipr.2016.0515.

[26] R. Rajeev, J. Abdul Samath, and N. K. Karthikeyan. "An Intelligent Recurrent Neural Network with Long Short-Term Memory (LSTM) BASED Batch Normalization for Medical Image Denoising." *Journal of medical systems,* Vol. 43, No. 8 , June 2019,pp 234. https://doi.org/10.1007/s10916-019-1371-9.

[27] M.F. Kazemi, M. A. Pourmina, and A. H. Mazinan. "Analysis of watermarking framework for color image through a neural network-based approach." *Complex & Intelligent Systems,* Vol.6, January2020, pp.213-220. https://doi.org/10.1007/s40747-020-00129-4.

[28] N. Leite, F. Melício, and A.C. Rosa. "A fast simulated annealing algorithm for the examination timetabling problem." *Expert Systems with Applications,* Vol122, May2019, pp. 137-151. DOI:https://doi.org/10.1016/j.eswa.2018.12.048.

[29] S. Kaur, L.K. Awasthi, A. L. Sangal, and G. Dhiman. "Tunicate Swarm Algorithm: A new bio-inspired based metaheuristic paradigm for global optimization." *Engineering Applications of Artificial Intelligence* Vol. 90, pp. 103541. https://doi.org/10.1016/j.engappai.2020.103541.

[30] W. Ding, Y. Ming, Z. Cao, and C.T. Lin. "A Generalized Deep Neural Network Approach for Digital Watermarking Analysis. "*IEEE Transactions on Emerging Topics in Computational Intelligence.* 2021.

[31] S. Hochreiter and J. Schmidhuber. "Long short-term memory". *Neural computation*, Vol.9, no. 8, 1997, pp. 1735-1780.

[32] H.W. Makram, J-F. Couchot, R. Couturier, and R. Darazi. "Using Deep learning for image watermarking attack." *Signal Processing: Image Communication*, Vol. 90, October2020, pp.116019. https://doi.org/10.1016/j.image.2020.116019.

R. Radha Kumari

R. Radha Kumari working as Associate Professor in SACET (Affiliated to JNT University, Kakinada), Chirala, India, and she has 17 years of teaching experience. She has received M.Tech. Degree in VLSI Design from Sathyabama university, Chennai and B.Tech in Electronics & Communication Engineering from N.B.K.R.I.S.T, Vidyanagar, Nellore(dt) A.P. Her areas of interest includes VLSI Design, Digital image processing and Artificial Intelligence, Machine Learning.

Dr. V. Vijaya Kumar

Dr. V. Vijaya Kumar is working as Professor & Dean Department of Computer Science & Engineering and Information technology in Anurag Group of Institutions (Autonomous) Hyderabad. He received integrated M.S.Engg, in CSE from USSR in 1989. He received his Ph.D. from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India in 1998 in CSE and guided 33 research scholars for Ph.D. He acted as principle investigator for various R&D projects. He has served JNT University for 13 years as Assistant Professor, Associate Professor and Professor. He has received Distinguished Professor award from Computer Science of India (CSI), Mumbai, best researcher and best teacher award from JNT University, Kakinada, India, Leading Scientist of the WORLD -2009 and Top 100 Scientists award in 2010 from International

Biographical Centre, Cambridge, England. His research interests include Big data and image analytics, image retrieval, texture analysis, author attribution, digital water marking. At present he is also acting as BoS member for various universities and institutions. He is the life member of CSI, ISCA, ISTE, IE (I), IETE, ACCS, CRSI, IRS and REDCROSS. He published more than 120 research publications till now in various national, international journals and conferences. He has delivered key note addresses at various international conferences.

Dr. K. Rama Naidu

Dr. K. Rama Naidu obtained his B.Tech (ECE) from JNTU College of Engineering, Anantapur in 1983 and M.Tech (Microwave Engineering) from Institute of Technology, Banaras Hindu University, 1990. He obtained his Ph.D. (Wireless Communications) from Indian Institute of Technology, Kharagpur in 2009. His areas of research are channel modelling, resource allocation for wireless relay systems, PAPR reduction in OFDM, channel estimation, and cognitive radios. He has published and presented more than 28 technical papers in various International Journals & conferences. He has 29 years of experience in teaching. Currently he is working as Professor, Dept. of ECE, at JNTUA College of Engineering, Ananthapuramu, A. P.

A Study on RGB Image Multi-Thresholding using Kapur/Tsallis Entropy and Moth-Flame Algorithm

V. Rajinikanth[1], Seifedine Kadry[2], Rubén González Crespo[3]*, Elena Verdú[3]

[1] Department of Electronics and Instrumentation Engineering, St. Joseph's College of Engineering, Chennai 600119, TN (India)
[2] Faculty of Applied Computing and Technology, Noroff University College, Kristiansand (Norway)
[3] School of Engineering and Technology, Universidad Internacional de la Rioja (UNIR), Logroño (Spain)

Received 16 May 2021 | Accepted 18 November 2021 | Published 24 November 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

In the literature, a considerable number of image processing and evaluation procedures are proposed and implemented in various domains due to their practical importance. Thresholding is one of the pre-processing techniques, widely implemented to enhance the information in a class of gray/RGB class pictures. The thresholding helps to enhance the image by grouping the similar pixels based on the chosen thresholds. In this research, an entropy assisted threshold is implemented for the benchmark RGB images. The aim of this work is to examine the thresholding performance of well-known entropy functions, such as Kapur's and Tsallis for a chosen image threshold. This work employs a Moth-Flame-Optimization (MFO) algorithm to support the automatic identification of the finest threshold (Th) on the benchmark RGB image for a chosen threshold value (Th=2,3,4,5). After getting the threshold image, a comparison is performed against its original picture and the necessary Picture-Quality-Values (PQV) is computed to confirm the merit of the proposed work. The experimental investigation is demonstrated using benchmark images with various dimensions and the outcome of this study confirms that the MFO helps to get a satisfactory result compared to the other heuristic algorithms considered in this study.

Keywords

Finest Threshold, Kapur's Entropy, Tsallis Entropy, Moth-Flame-Optimization, Picture Quality, PSNR, SSIM.

DOI: 10.9781/ijimai.2021.11.008

I. Introduction

Bi-level and multi-level image thresholding is widely employed to improve the quality of the image by grouping the pixels based on the selected threshold level. In literature, a number of image threshold methods are implemented on a class of gray/RGB scale images due to their practical significance. Recently, threshold methods are employed to pre-process a range of images with varied dimension and different pixel distributions [1], [2]. More commonly, the images, such as benchmark photographs [3]-[5], satellite images [6] and medical pictures [7] are thresholded with various methodologies and different optimization algorithms. A considerable number of thresholding methods, such as Otsu, Kapur, Tsallis, Renyi and Shannon are already implemented on a variety of digital images and every work discussed its own contribution. The review of the existing threshold identification can be found in [8], [9].

The concept used in the threshold operation is; adjusting the thresholds of the given image to enhance its features by grouping the similar pixels based on the chosen threshold value. This process is normally carried out with the help of computer algorithms and this procedure is terminated based on a chosen Objective Function (OF). In most of the cases, maximization of OF is preferred and the chosen algorithm will continuously work to maximize the OF by adjusting the thresholds arbitrarily.

In most of the medical image processing techniques, the thresholding is chosen as a pre-processing procedure and the outcome in medical data assessment depends on the implemented threshold operation [10]. Further, entropy supported functions are widely used in medical data assessment [11]-[13] and hence, it is essential to identify the suitable entropy based technique to pre-process the gray/RGB scale picture [14]. Assessment of RGB scaled image is quite complex due to its complex histogram and the methodology which works well on RGB scale images can be easily transferred to pre-process the gray scale images. Further, the complexity of the image processing task also will increase based on the dimension of the image and hence, in this work the benchmark images with dimensions; 512 × 512 × 3 and 720 × 576 × 3 are considered for the demonstration.

The proposed work aims to evaluate the threshold performance of Kapur's/Tsallis entropy functions on a chosen RGB test picture. Both these entropy functions work by identifying and improving the essential pixels of the image which consist the key pixel groups. Threshold identification in RGB grade image is quite complex compared to the gray scale image, hence, this work employed the traditional Moth-Flame-Optimization (MFO) algorithm and the performance of the proposed technique is verified based on the computed Picture-Quality-Values (PQV). The experimental outcome of this study confirms that the PQV achieved for Tsallis is better in RGB scaled images compared

* Corresponding author.

E-mail address: ruben.gonzalez@unir.net

to Kapur. This result confirms that the RGB scale images thresholded with Tsallis helps to achieve a better pixel grouping and this procedure can be considered to examine the traditional and medical grade digital pictures to get better results during the examination.

The proposed work considered the following procedures to improve the outcome of thresholding:

- Implementing the bounded search procedure discussed by Raja et al. [15] to minimize the search time.
- Implementing the modified objective function discussed in Rajinikanth and Couceiro [16] to enhance the outcome.
- Experimental investigation is executed on a commonly used RGB scale benchmark images with two different dimensions.

II. Related Earlier Works

Image multi-thresholding is one of the common techniques to enhance the test image with a chosen procedure. The entropy based methods are normally considered in the literature to enhance the vital information in the chosen test picture. When a medical image (Gray/RGB) is to be processed, the entropy based thresholding helps to provide better pixel grouping, which improves the visibility of the abnormal region of the image, which is to be examined. Implementation of entropy supported medical image enhancement is common procedure and Kapur's/Tsallis thresholding schemes are widely employed in this task. The Kapur's/Tsallis based medical image thresholding can be found in the earlier research work [10].

Examination of gray scale picture is quite simple and a manual threshold selection procedure can also be employed to improve the quality of the gray scale picture. The assessment of the RGB scaled image is one of the complex tasks due to its complex histogram and hence, a considerable number of procedures are developed to evaluate the RGB scale pictures.

Table I summarizes few recently implemented entropy based thresholding techniques employed to threshold the RGB images. The considered RGB image thresholding procedure helps to improve the information in the chosen test picture by grouping the similar pictures and most of the recently developed RGB thresholding procedures employed a chosen heuristic algorithm to identify the threshold value automatically by maximizing the entropy value.

The information presented in Table I confirms that the multi-level thresholding with entropy functions are a widely adopted procedure, which provides a significant result on a class of traditional and medical grade images. The entropy supported medical image examination is a widely adopted pre-processing technique in which the thresholding helps to enhance the abnormal/disease section in the image, which is then extracted and assessed with a chosen segmentation method. The multi-level thresholding with Th=2,3,4,5 is a commonly adopted method and the results of earlier works presented in Table I confirms that the entropy supported thresholding helps to get a better PQV compared to Otsu's between class variance technique. Hence, this work employed the entropy supported thresholding to enhance the RGB scaled pictures with varied dimensions.

III. Methodology

The heuristic algorithm based threshold identification has largely attracted the research community to pre-process the gray/RGB scale images, due to its wide use. The methodology, which works well on a class of RGB grade picture, will work on a class of gray scale pictures. The threshold methodology implemented in this work is depicted in Fig. 1 and this work considered the RGB pictures of dimension 512 × 512 × 3 and 720 × 576 × 3 pixels.

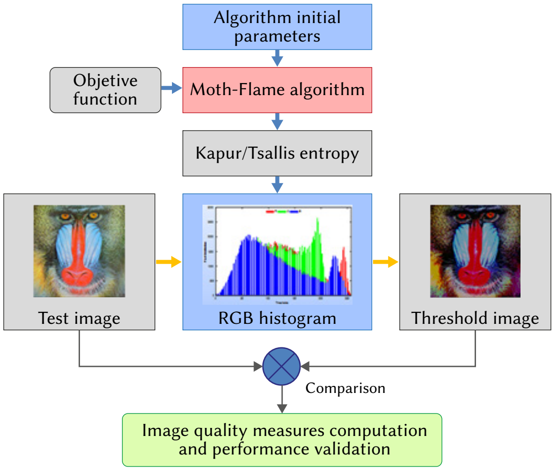


Fig. 1. Outline of methodology executed in the proposed work.

TABLE I. Summary of the RGB Scaled Image Thresholding With Entropy Function and Heuristic Algorithm

Reference	Methodology
Aziz et al. [1]	Multi-thresholding using Wale algorithm and MFO is implemented to find the optimal thresholds for a class of benchmark gray scale images.
Jia et al. [2]	Implementation of RGB image thresholding is presented using modified MFO.
Sathya et al. [17]	Implementation of RGB image thresholding using Kapur, Otsu and minimum cross entropy is presented using heuristic technique.
Farshi and Ardabili [18]	Implementation of multi-level thresholding using hybrid algorithm is presented using benchmark pictures.
Anitha et al. [19]	RGB image thresholding with modified whale optimization is presented for benchmark test pictures.
Kurban et al. [20]	A detailed review on RGB image thresholding for aerial images.
Bhandari [21]	Beta differential evolution algorithm supported fast RGB image thresholding is presented using benchmark pictures.
Xing [22]	RGB image multi-thresholding with improved emperor penguin optimization is presented.
Meyyappan et al. [23]	Skin melanoma image thresholding using Kapur's entropy and harmony search algorithm is demonstrated.
Borjigin and Sahoo [24]	RGNB image thresholding with Tsallis entropy is discussed.
Kadry and Rajinikanth [25]	Implementation of Tsallis entropy and MFO is presented for gray scale image thresholding.
Elaziz et al. [26]	Multilevel thresholding of benchmark pictures are presented using whale optimization algorithm.
Elaziz et al. [27]	Multi-level thresholding with Harris hawks algorithm is demonstrated.

Initially, a chosen threshold method with a chosen optimization algorithm is implemented to enhance the considered test picture. The implemented method will randomly adjust the threshold of the picture, till the objective function value is maximized or a maximum number of iterations is reached. After getting the pre-processed image, the performance of this image is then computed based on a comparative analysis with the original picture. During this process, the essential PQVs are computed and, based on these values; the performance of the implemented threshold operation is confirmed.

Executing the multi-thresholding on a chosen test picture requires the following procedures:

- Choice of appropriate thresholding methodology.
- Choice of the heuristic technique.
- Selection of the objective function.
- Selection of PQV to assess the outcome of the experiment.

A. Entropy Function

Entropy based methods are largely used in image as well as signal evaluation methods to identify the abnormal regions. The earlier research works confirmed that the entropy based procedures offers better results compared to other similar methodologies.

In image examination applications, Kapur's Entropy (KE) and Tsallis Entropy (TE) are widely employed in situations when the key pixel groups in the image are to be enhanced based on a chosen threshold [10]. The processing methodology related to the KE is simple and it can be mathematically denoted using a simple probability distribution function. The execution steps in TE are quite complex and its probability function is monitored using an entropic index. This research work aims to evaluate the thresholding performance of KE and TE on chosen benchmark RGB scale images.

1. Kapur's Entropy

KE was initially proposed by Kapur et al. [28] to pre-process a class of gray scale images. In this work, the histogram values are randomly adjusted till the entropy of the histogram reaches a maximized value.

The KE for a gray scale image is mathematically described as follows:

Let, $Th = [T_1, T_2, ..., T_{n-1}]$ denote the threshold vector for gray picture and for the case of the RGB, it can be represented as; $Th_R = [T_{R1}, T_{R2}, ..., T_{Rn-1}]$, $Th_G = [T_{G1}, T_{G2}, ..., T_{Gn-1}]$ and $Th_B = [T_{B1}, T_{B2}, ..., T_{Bn-1}]$.

The alteration in threshold is done, till the following condition is maximized:

$$J_{KE_{max}} = F(Th) = \sum_{j=1}^{n} H_j^C \tag{1}$$

where H_j^C = probability distribution and C=image class identification (C = 1 for gray image and C = 3 for RGB).

The KE described in this section can be adopted for the RGB scale images, in which the R,G and B histogram is separately determined with the chosen function and this evaluation is continued till the average entropy value for the image is maximized ($J_{KE_{max}}$). Other related information for the KE can be found in [2], [10].

2. Tsallis Entropy

The idea of TE was derived from Shannon's Function (SF) depicted in Eqn. (2) and this equation forms the SF when $\varepsilon \to 1$.

$$SF = \frac{1-\sum_{j=1}^{Th}(p_j)^{\varepsilon}}{\varepsilon-1} \tag{2}$$

where Th = total thresholds and ε = entropy indicator.

The simulated additively rule for the entropy is as follows:

For a gray scale picture with a threshold of range [0, 1, ..., L − 1] along with probability distributions $P_i = P_0, P_1, ..., P_{L-1}$, it is considered to identify the final entropy function based on assigned threshold.

For the RGB class image, this distribution is as follows; $P_{Ri} = P_{R0}, P_{R1}, ..., P_{RL-1}, P_{Gi} = P_{G0}, P_{G1}, ..., P_{GL-1}$ and $P_{Bi} = P_{B0}, P_{B1}, ..., P_{BL-1}$.

Compared to the KE, the execution of the TE is quite complex, since, its outcome depends mainly on the probability distributions, which decide the maximized entropy value. Other related details of TE can be found in the earlier works [25], [29].

B. Moth-Flame-Optimization

In the literature, a number of Heuristic Algorithms (HA) are proposed and implemented by the researchers to find the optimal solution for a class of real world problems. The performance of a chosen HA depends on its updating mechanism and in most of the existing algorithms, the movement of agents from the current location to the new location happens randomly. For a chosen problem, the search of the optimal solution depends on the dimension of the solution and the search methodology which moves the agents towards the optimal solution. Most of the existing HA considered random search processes to move the agents towards the optimal location. Further, the search procedures, such as chaotic search, Lévy-flight, and Brownian-distribution are also adopted to move the agents towards the finest solution with lesser iteration. Recently, to move the search agents quickly towards the optimal solution, the spiral and spherical search strategies are proposed to get the optimal solution to a chosen problem [1], [2], [25].

MFO is a nature inspired HA invented by Mirjalili in 2015 [30] to find the optimal solution for numerical benchmark problems. The concept of MFO is based on the movement of a Moth towards Flame based on a pre-defined pattern (spiral). In this algorithm, the moths are the search agents and the flame is the solution for the problem. If the algorithm search is initiated with a number of agents (moths), then every agent is allowed to reach their associated flame (solution) using a predefined defined search-pattern. The conventional MFO with a single agent searching the solution in a 3D space is depicted in Fig. 2. All the agents are randomly initiated in the search universe based on the dimension of the search and every agent is allowed to converge towards the solution when the search iteration increases. The main merit of this algorithm is that every agent travels in a spiral shaped search path, which helps to reduce the number of iterations to find the optimal solution compared to the random, Lévy-flight, and Brownian-distribution functions. Further, the search pattern of MFO is predefined and it does not take the arbitrary path to reach the solution during an optimization task and this procedure makes the MFO more successful in finding the solutions for a chosen problem compared to the existing algorithms in the literature.

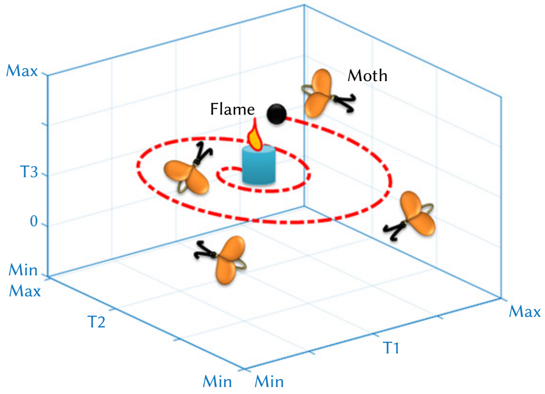


Fig. 2. Search pattern of a Moth towards a Flame.

TABLE II. Heuristic Algorithms and their Parameters Considered in This Work to Justify the Performance of MFO

Measures	PSO [6]	BA [4]	FA [16]	MA [37]	AOA [38]	MFO
Number of agents	20	20	20	20	20	20
Search dimension	Th=2,3,4,5	Th=2,3,4,5	Th=2,3,4,5	Th=2,3,4,5	Th=2,3,4,5	Th=2,3,4,5
Search pattern	Random search	Ikeda-Map	Lévy-flight	Lévy-flight	Multiple search pattern	Spiral combined with random search
Objective function	MOF	MOF	MOF	MOF	MOF	MFO
Maximum Iteration ($Iter_{max}$)	2000	2000	2000	2000	2000	2000
Search termination	$Iter_{max}$	$Iter_{max}$	$Iter_{max}$	$Iter_{max}$	$Iter_{max}$	$Iter_{max}$

To define the mathematical model for MFO; let us consider that a chosen search space has a-number of moths (M) and b-number of flames (F). Let the initial distance among M and F is $D_a = |F_b - M_a|$, which is to be minimized to find the optimal solution.

The mathematical expression for a moth progress towards the flame can be defined as:

$$M_a = D_a * e^{\kappa\Re} * cos(2\pi\Re) + F_b \tag{3}$$

where M_a = a^th moth, F_b = b^th flame, κ = constant to define spiral pattern, and $\Re$ = random variable of range [-1,1].

In this equation, the random variable $\Re$ can be used to speed up the spiral shape formation process to guide the Moths towards the flame. The complete information about the traditional and improved MFO can be found in [1], [2], [25], [31]-[33].

C. Implementation

In this work, the MFO algorithm is considered to identify the optimal threshold for the RGB scaled images using the KE/TE based on a chosen objective function. Initially, the test image is pre-processed based on a chosen threshold (Th = 2, 3, 4, 5) by maximizing the entropy value. After thresholding the image based on a chosen Th, a pixel wise comparison is executed to confirm the enhancement in the thresholded image.

In the literature, multi-level thresholding with Single-Objective-Function (SOF) and weighted sum of Multiple-Objective-Function (MOF) is very common and the earlier work confirmed that the MOF helps to get better enhancement compared to the SOF. Hence, this work employed the MOF to threshold the RGB scale image with KE/TE.

The MOF considered in this work is depicted in Eqn. (4):

$$MOF_{max} = (W_1 * Entropy) + (W_2 * PSNR) + (W_3 * SSIM) \tag{4}$$

where, Entropy = maximized KE/TE, PSNR = Peak signal-to-noise ratio, SSIM = Structural Similarity Index $W_1 = 1$, $W_2 = 0.5$, and $W_3 = 0.5$.

In this work, HA search with KE/TE is performed to satisfy Eqn. (4).

D. Picture-Quality-Value Computation

The performance of the image processing scheme is confirmed by computing the image related measures. From the image thresholding literature, it is noted that the merit of the image thresholding process can be confirmed by computing the necessary PQV. This PQV can be obtained by comparing the original and threshold image.

The quality of the thresholding can be confirmed by computing the measures, such as Mean Squared Error (MSE), Root MSE (RMSE), PSNR, Average Difference (AD), Structural Content (SC), Normalized Absolute Error (NAE) and SSIM.

All these measures help to confirm that the thresholded image is having better enhancement compared to the original test picture. MSE and RMSE are the measures of the intensity variation in the processed image compared to the original picture. The PSNR provides the ratio among the highest potential pixel of an image and the power of distorting noise which affects the quality of threshold image. The AD presents the possible deviation among the original and processed image. SC confirms the presence of the vital information in the threshold image compared to the original image. NAE is the difference between the original and processed picture and SSIM presents the image quality degradation due to thresholding.

Let us consider that O and P denote the dimension of real (R) and threshold (T) pictures, and the mathematical expression of these measures can be expressed as:

$$\text{MSE} = \frac{1}{OP}\sum_{j=1}^{O}\sum_{n=1}^{P}(R_{j,n} - T_{j,n}) \tag{5}$$

$$\text{RMSE} = \sqrt{MSE} \tag{6}$$

$$\text{PSNR} = 10\, log\frac{(255)^2}{MSE} \tag{7}$$

$$\text{AD} = \frac{\sum_{j=1}^{O}\sum_{n=1}^{P}(R_{j,n} - T_{j,n})}{OP} \tag{8}$$

$$\text{SC} = \frac{\sum_{j=1}^{O}\sum_{n=1}^{P}R_{j,n}}{\sum_{j=1}^{O}\sum_{n=1}^{P}T_{j,n}} \tag{9}$$

$$\text{NSE} = \sum_{j=1}^{O}\sum_{n=1}^{P}(R_{j,n} - T_{j,n}) \tag{10}$$

$$\text{SSIM} = \frac{(2\mu_o\mu_p + G_1)(2\sigma_{op} + G_2)}{(\mu_o{}^2 + \mu_p{}^2 + G_1)(\sigma_o{}^2 + \sigma_p{}^2 + G_2)} \tag{11}$$

where, μ_o and μ_p are mean values, σ_o^2 and σ_p^2 variances and σ_{op} correlation coefficient. Further, $G_1 = (0.01L)^2$ and $G_2 = (0.03L)^2$.

Other information on these PQVs can be found in the literature [34]-[36].

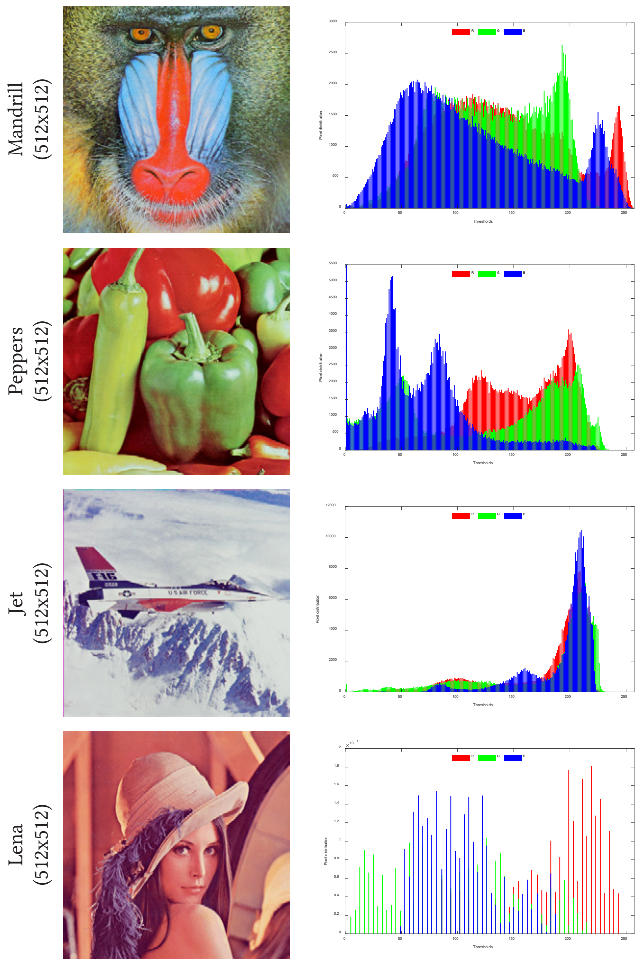
E. Performance Evaluation

The performance of the proposed thresholding by MOA based KE/TE is then validated using the well known HA, such as Particle Swarm Optimization (PSO), Bat Algorithm (BA), Firefly Algorithm (FA), Mayfly Algorithm (MA) and Aquila Optimization algorithm (AOA). To have a fair evaluation, every algorithm is assigned with similar agents, search dimension, objective function, $Iter_{max}$ and termination as depicted in Table II. Every algorithm has its own search pattern which influences the search convergence and the attained image quality. Compared to MA and AOA, the implementation steps involved in MFO are quite simple, Further, the PSO, BA and FA follow a complex search pattern and hence the results by the MFO are satisfactory on the chosen test images.

IV. Results and Discussions

This work aims to demonstrate a multi-threshold scheme for a class of images with varied dimensions. The experimental investigation is performed using MATLAB software and the proposed technique is independently tested on all the considered imagery with a threshold value ranging from 2 to 5. This work executed an entropy based methodology to find the finest threshold with the help of MFO algorithm.

Later an assessment of the threshold image and original picture is performed to find the PQVs and based on these values; the performance of this system is validated. Primarily, this work is tested on the benchmark RGB pictures shown in Fig. 3. This figure presents the chosen trial images (Fig. 3(a)) along with its histogram (Fig. 3(b)). The histogram of RGB images is very complex and hence, identification of the finest thresholds is quite difficult. Hence, this work implemented a bounded threshold technique discussed by Raja et al. [15]. In bounded search, instead of keeping $Th_{min} = 0$ and $Th_{max} = 255 = L-1$, a boundary is assigned based on its pixel strength.



(a) (b)

Fig. 3. Chosen digital photographs of dimension 512x512x3.

The bounded search is executed separately on R, G and B thresholds as follows:

$$R_{channel} = Th_{min} < R < Th_{max}$$

$$G_{channel} = Th_{min} < G < Th_{max}$$

$$B_{channel} = Th_{min} < B < Th_{max}$$

If the minimum and maximum thresholds are assigned for each channel, then it is easy for the MFO to find the finest threshold with minimal iteration.

Fig. 4 depicts the R,G,B histogram for Mandrill image, for which the threshold search boundary is assigned as follows:

$$R_{channel} = 28 < R < 248$$

$$G_{channel} = 20 < G < 216$$

$$B_{channel} = 8 < B < 245$$

The bounded search helped to achieve satisfactory results with lesser search iteration. In this work a ten-fold cross validation is implemented for every image with every threshold and the best result among them is considered as the optimized threshold.

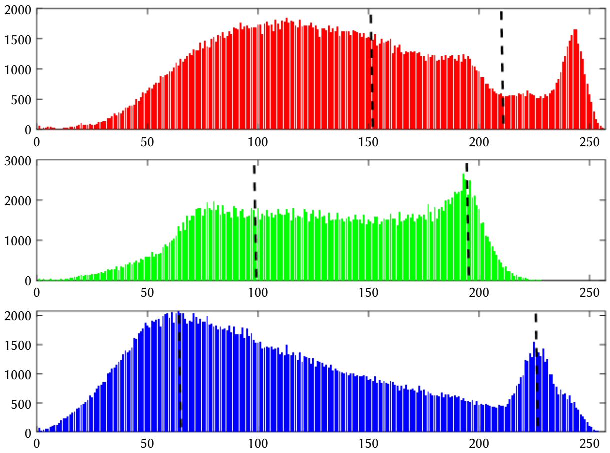


Fig. 4. Bi-level threshold implemented on RGB scale histogram.

Initially, the KE based multilevel thresholding is implemented using Mandrill picture for a chosen Th of 2 to 5. For this image, every algorithm helped to identify similar objective value with approximately similar PQV. But the MFO showed a steady search behavior due to its spiral search operator compared to other algorithms. The search convergence by the FA and MA is better compared to the MFO due to its Lévy-flight process and compared to PSO, BA and AOA, the search convergence of MFO is better. The performance measures computed for the Mandrill with KE and Th=5 is presented in Fig. 5, which confirms that the overall PQV obtained by the MFO is satisfactory compared to other HA. Fig. 6 presents a performance comparison of MFO with other methods and this comparison confirms that the MFO's performance is better compared to PSO, BA and FA and approximately similar to MA and AOA.

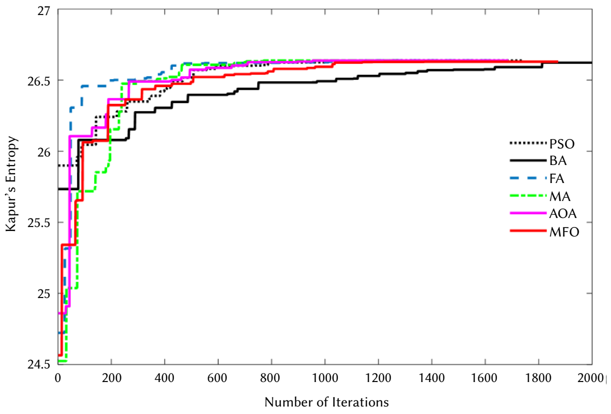


Fig. 5. Search convergence of chosen HA for Mandrill with KE and Th=5.

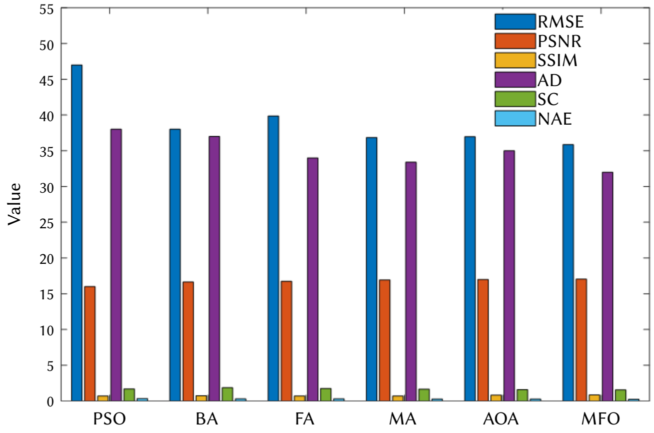


Fig. 6. Performance evaluation of MFO with other HA.

The results attained with the MFO and the KE (MFO+KE) are depicted in Fig. 7 for the chosen thresholds. Fig. 7(a) to (d) depict the results attained for Th=2 to 5, respectively. After recording the thresholded image, an assessment of this image with its original image is performed to compute all the possible PQVs discussed in sub-section III.D.

A pixel wise comparison among the images is performed to compute the essential PQVs for each image. Initially, this computation is implemented for the Mandrill image and the attained SSIM map is depicted in Fig. 8. Other PQVs attained with this procedure are presented in Table III. The values of this table confirm that, the lesser threshold gives lower values of the PQVs and this value gradually improves for higher values of threshold.

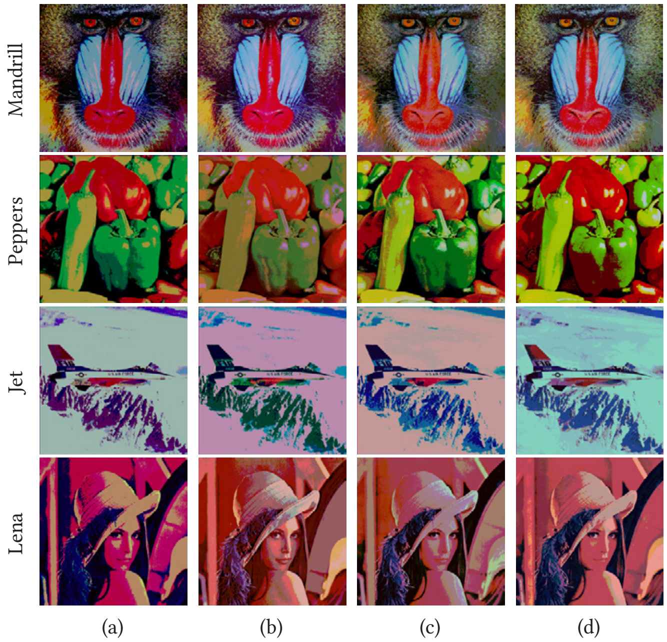


Fig. 7. Thresholding outcome with MFO+KE.

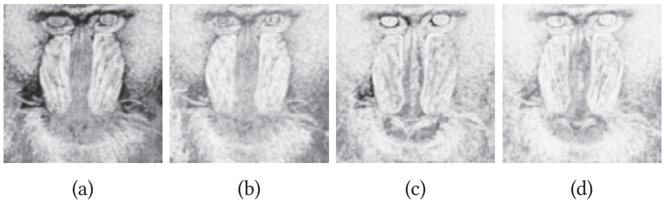


Fig. 8. SSIM map computed for Mandrill image.

TABLE III. PQVs Attained With MFO+KE

Image	Th	RMSE	PSNR	SSIM	AD	SC	NAE
Mandrill	2	63.9934	12.0081	0.6168	58.7727	2.2679	0.4534
	3	46.4955	14.7826	0.7614	40.1494	1.6591	0.3098
	4	41.4216	15.7863	0.7897	37.7085	1.7388	0.2909
	5	35.8594	17.0387	0.8427	31.9953	1.5582	0.2469
Peppers	2	51.4619	13.9011	0.6193	48.2072	2.2920	0.4010
	3	44.5729	15.1494	0.6700	38.7395	2.1937	0.3222
	4	43.8707	15.2873	0.6878	37.3242	1.6947	0.3105
	5	40.5123	15.9791	0.7889	34.9152	1.6760	0.2904
Jet	2	54.5974	13.3874	0.8085	52.8337	1.8570	0.2949
	3	45.6481	14.9424	0.8247	43.7349	1.6415	0.2441
	4	44.3914	15.1848	0.8555	40.2827	1.5176	0.2248
	5	33.4113	17.6529	0.9075	31.9359	1.4242	0.1782
Lena	2	64.6375	11.9211	0.4769	60.4672	2.6461	0.5021
	3	36.7225	16.8322	0.7460	34.1909	1.7332	0.2839
	4	35.8501	17.0410	0.7827	32.9572	1.6923	0.2737
	5	33.7777	17.5582	0.7891	31.0726	1.6477	0.2580

A similar procedure is then repeated using the TE (MFO+TE) on the considered test images and the corresponding results are depicted in Fig. 9 and Table IV. Fig. 9(a) to (d) presents the results when the chosen thresholds are from 2 to 5.

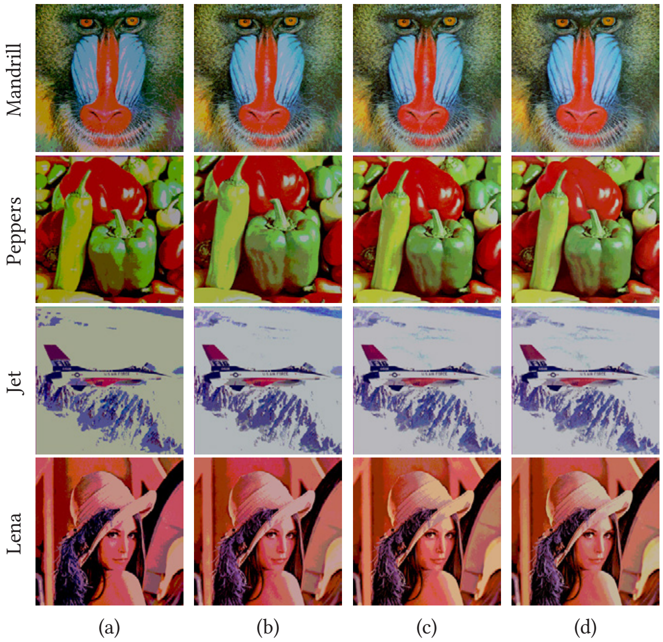
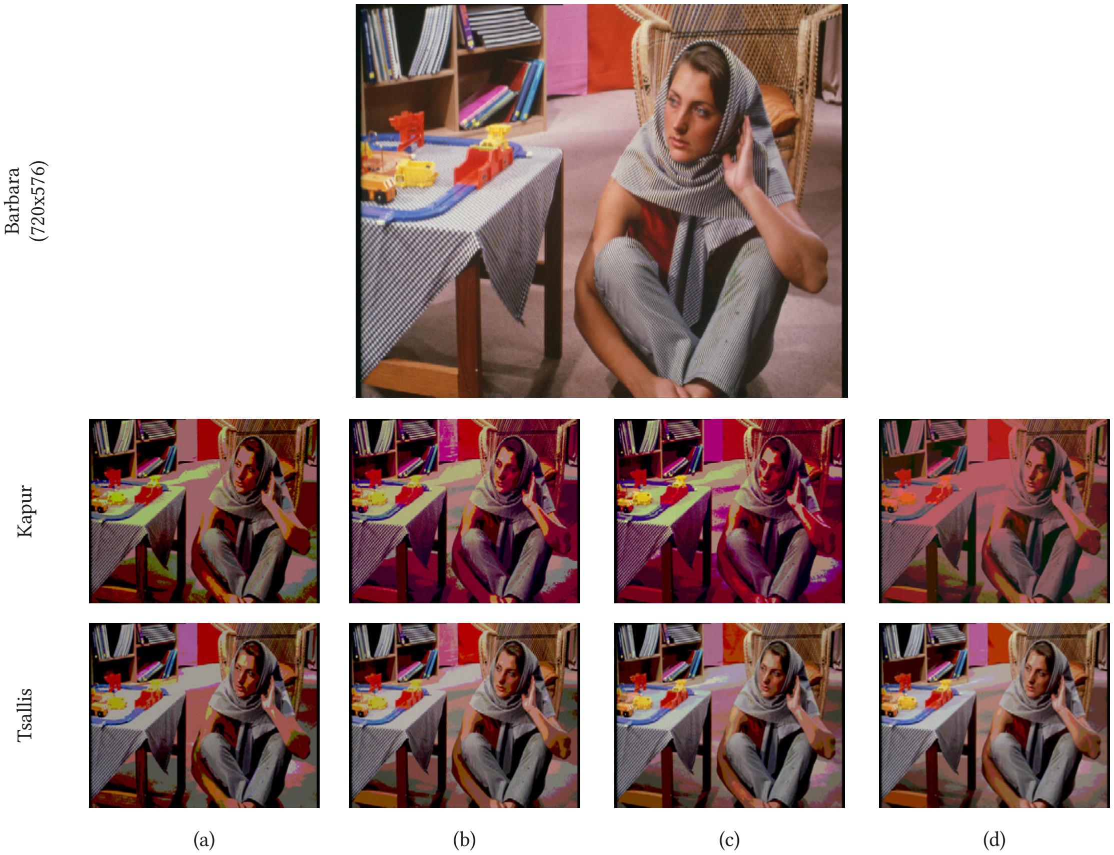


Fig. 9. Thresholding outcome with MFO+TE.

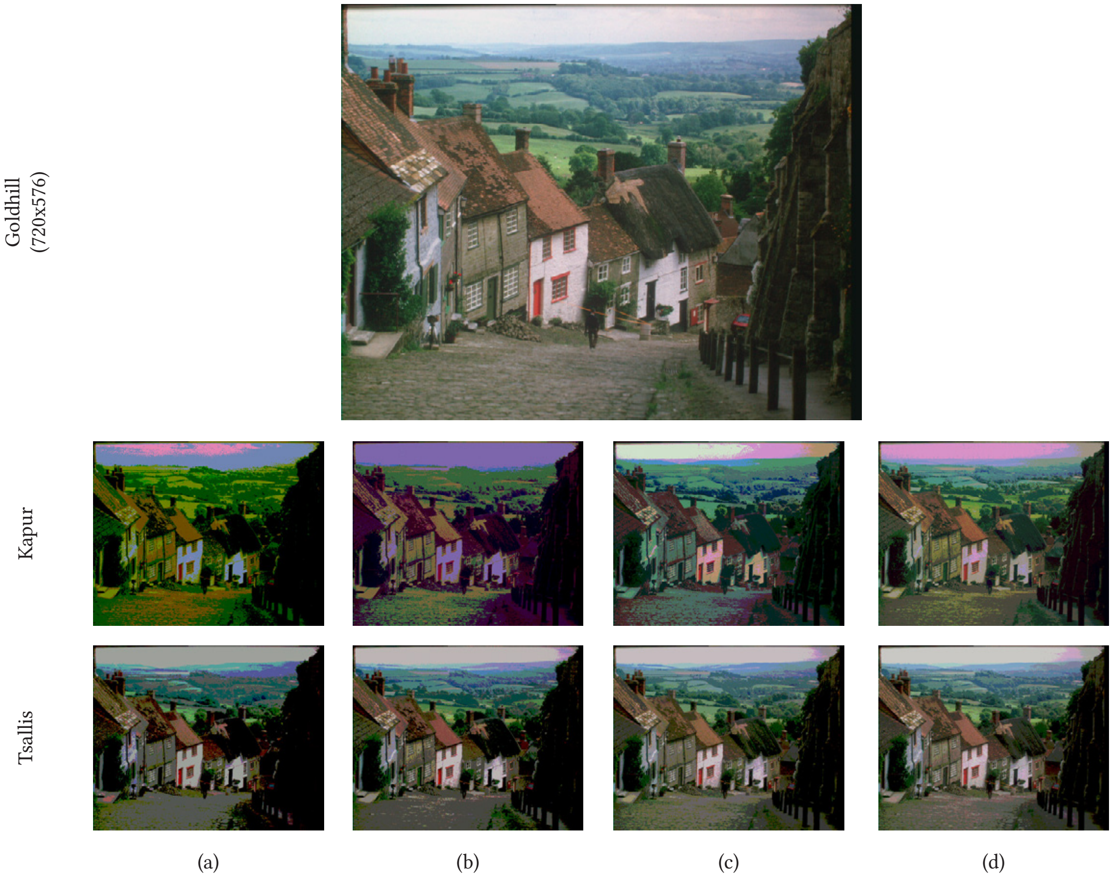
Finally, an analysis between Table III and Table IV is then executed to assess the PQVs attained with MFO+KE and MFO+TE and this comparison confirms that, the results attained with TE are superior for all the thresholds compared to the KE. Even though the PQVs of KE are lesser, the pixel grouping and the image enhancement achieved with the KE is superior compared to the TE.

Hence, during the real image examination tasks, the choice of a particular technique can be done based on PQVs or based on image enhancement. During the real image processing task, the choice and implementation of a particular thresholding procedure depends on the operator and its expertise.



(a) (b) (c) (d)

Fig. 10. Thresholding results for Barbara image.



(a) (b) (c) (d)

Fig. 11. Thresholding results for Goldhill picture.

TABLE IV. PQVs Computed for the Outcome of MFO+TE

Image	Th	RMSE	PSNR	SSIM	AD	SC	NAE
Mandrill	2	42.9311	15.4754	0.8161	40.8138	1.9045	0.3149
	3	30.7935	18.3616	0.8955	28.9921	1.5468	0.2237
	4	24.8261	20.2326	0.9308	23.2634	1.4127	0.1795
	5	20.7107	21.8069	0.9496	19.3494	1.3278	0.1493
Peppers	2	40.0633	16.0759	0.7543	37.6425	1.8912	0.3131
	3	30.6295	18.4080	0.8263	28.0955	1.5433	0.2337
	4	25.1771	20.1107	0.8568	22.8180	1.3914	0.1898
	5	18.8785	22.6115	0.9069	17.2850	1.2982	0.1438
Jet	2	43.8722	15.2870	0.6715	41.6883	1.6468	0.2327
	3	29.2309	18.8140	0.7817	27.2913	1.3549	0.1523
	4	24.4745	20.3565	0.8020	22.8037	1.2900	0.1273
	5	21.2613	21.5790	0.8194	19.6579	1.2465	0.1097
Lena	2	41.1201	15.8497	0.6715	37.9305	1.7185	0.3150
	3	31.1058	18.2740	0.7817	28.8739	1.5080	0.2398
	4	27.1371	19.4595	0.8020	24.4986	1.3750	0.2034
	5	26.5207	19.6591	0.8194	23.9857	1.3641	0.1992

This work demonstrated a satisfactory result with the KE and better result with the TE with respect to the PQVs on image with dimension 512x512x3 pixels and in future, the proposed approach can be used to test other existing benchmark gray/RGB class pictures.

A similar procedure is then executed for the test images with dimension 720x576x3 pixels and the corresponding outcomes are depicted in Fig. 10 and Fig. 11 for various thresholds with chosen entropy values.

The results of the proposed study confirm that, the proposed thresholding work helps to attain better results on a class of RGB images. Further, the proposed work confirms that the overall performance of the TE based thresholding on RGB image is better compared to KE.

The future scope of this work is as follows:

- The proposed work can be employed to threshold the gray scale images.
- The MOF value can be enhanced by including more PQVs.
- This work currently implemented a traditional MFO algorithm which works based on a random operator, $\Re$ whose value varies with a range [-1,1]. In future, this random variable can be replaced with various search operators (Ex. Levy, Brownian-distribution and chaotic operator) to enhance the convergence of MFO.

V. Conclusion

This research aims to recommend a methodology to solve the multi-thresholding problem of RGB scale images using entropy value. This scheme used the thresholds ranging from 2 to 5 and to realize the optimum threshold, MFO is employed. This work proposes a random search along with a novel Multiple-Objective-Function (MOF). The role of MFO is to randomly adjust the thresholds till the MOF is maximized. This work is separately tested with KE and TE techniques on the chosen benchmark images. After discovering the necessary threshold for the chosen picture, in order to validate its performance, a comparison is performed among the original and threshold image to find the PQVs. Based on these values, the performance is confirmed. From the attained results, it is established that the proposed system works well on a class of RGB images with different dimensions. The experimental outcome confirms that the PQVs achieved with the TE are better compared to the KE. Further, a comparative assessment with KE confirmed that the performance of MFO is better compared to PSO, BA and FA and approximately similar to MA and AOA.

References

[1] M.A.E. Aziz, A.A. Ewees, A.E. Hassanien, “Whale Optimization Algorithm and Moth-Flame Optimization for multilevel thresholding image segmentation,” *Expert Systems with Applications*, vol. 83, pp. 242-256, 2017, https://doi.org/10.1016/j.eswa.2017.04.023.

[2] H. Jia, J. Ma, W. Song, “Multilevel Thresholding Segmentation for Color Image Using Modified Moth-Flame Optimization,” *IEEE Access*, vol. 7, pp. 44097- 44134, 2019, DOI: 10.1109/ACCESS.2019.2908718.

[3] V. Rajinikanth, N.S.M. Raja, S.C. Satapathy, “Robust color image multi-thresholding using between-class variance and cuckoo search algorithm,” *Advances in Intelligent Systems and Computing*, vol. 433, pp. 379-386, 2016, https://doi.org/10.1007/978-81-322-2755-7_40.

[4] S.C. Satapathy, N.S.M. Raja, V. Rajinikanth, A.S. Ashour, N. Dey, “Multi-level image thresholding using Otsu and chaotic bat algorithm,” *Neural Computing and Applications*, vol. 29, no. 12, pp. 1285-1307, 2018, https://doi.org/10.1007/s00521-016-2645-5.

[5] A. Bahriye, “A study on particle swarm optimization and artificial bee colony algorithms for multilevel thresholding,” *Applied Soft Computing*, vol. 13, no. 6, pp. 3066–3091, 2013, https://doi.org/10.1016/j.asoc.2012.03.072.

[6] P. Ghamisi, M.S. Couceiro, F.M.L. Martins, J.A. Benediktsson, “Multilevel image segmentation based on fractional-order Darwinian particle swarm optimization,” *IEEE Transactions on Geoscience and Remote sensing*, vol. 52, no.5, pp. 2382–2394, 2014.

[7] S.L. Fernandes, V. Rajinikanth, S. Kadry, “A hybrid framework to evaluate breast abnormality using infrared thermal images,” *IEEE Consumer Electronics Magazine*, vol. 8, no. 5, pp. 31-36, 2019, doi: 10.1109/MCE.2019.2923926.

[8] M. Sezgin, B. Sankar, “Survey over Image Thresholding Techniques and Quantitative Performance Evaluation,” *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146– 165, 2004.

[9] M. Tuba, “Multilevel image thresholding by nature-inspired algorithms: A short review,” *Computer Science Journal of Moldova*, vol. 22, no. 3, pp. 318–338, 2014.

[10] V. Rajinikanth, S.C. Satapathy, S.L. Fernandes, S. Nachiappan, “Entropy based segmentation of tumor from brain MR images–a study with teaching learning based optimization,” *Pattern Recognition Letters*, vol. 94, pp. 87-95, 2017. https://doi.org/10.1016/j.patrec.2017.05.028.

[11] N. Arunkumar, K. Ramkumar, V. Venkatraman, E. Abdulhay, S.L. Fernandes, S. Kadry, S. Segal, “Classification of focal and non focal EEG using entropies,” *Pattern Recognition Letters*, vol. 94, pp. 112-117, 2017, https://doi.org/10.1016/j.patrec.2017.05.007.

[12] S.Z. Abbas, W.A. Khan, S. Kadry, M. Ijaz Khan, M. Waqas, M. Imran Khan, “Entropy optimized Darcy-Forchheimer nanofluid (Silicon dioxide, Molybdenum disulfide) subject to temperature dependent viscosity,” *Computer Methods and Programs in Biomedicine*, vol. 190, 105363, 2020, https://doi.org/10.1016/j.cmpb.2020.105363

[13] S.Z. Abbas, M. Ijaz Khan, S. Kadry, W.A. Khan, M. Israr-Ur-Rehman, M. Waqas, “Fully developed entropy optimized second order velocity slip MHD nanofluid flow with activation energy,” *Computer Methods and Programs in Biomedicine*, vol. 190, 105362, https://doi.org/10.1016/j.cmpb.2020.105362.

[14] S. Agrawal, R. Panda, S. Bhuyan, B.K. Panigrahi, “Tsallis entropy based optimal multilevel thresholding using cuckoo search algorithm,” *Swarm and Evolutionary Computation*, vol. 11, pp. 16–30, 2013.

[15] N.S.M. Raja, V. Rajinikanth, K. Latha, “Otsu based optimal multilevel image thresholding using firefly algorithm,” *Modelling and Simulation in Engineering*, vol. 2014, 794574, 2014, https://doi.org/10.1155/2014/794574.

[16] V. Rajinikanth and M.S. Couceiro, “Optimal multilevel image threshold selection using a novel objective function,” *Advances in Intelligent Systems and Computing*, vol. 340, pp. 177–186, 2015, https://doi.org/10.1007/978-81-322-2247-7_19.

[17] P. D. Sathya, R. Kalyani, V. P. Sakthivel, “Color image segmentation using Kapur, Otsu and Minimum Cross Entropy functions based on Exchange

Market Algorithm," *Expert Systems with Applications,* vol. 172, 114636, 2021.

[18] T. R. Farshi and A.K. Ardabili, "A hybrid firefly and particle swarm optimization algorithm applied to multilevel image thresholding," *Multimedia Systems*, vol. 27, no. 1, pp. 125-142, 2021.

[19] J. Anitha, S.I.A. Pandian, S.A. Agnes, "An efficient multilevel color image thresholding based on modified whale optimization algorithm," *Expert Systems with Applications*, vol. *178*, 115003, 2021.

[20] R. Kurban, A. Durmus, E. Karakose, "A comparison of novel metaheuristic algorithms on color aerial image multilevel thresholding," *Engineering Applications of Artificial Intelligence*, vol. *105*, 104410, 2021.

[21] A.K. Bhandari, "A novel beta differential evolution algorithm-based fast multilevel thresholding for color image segmentation," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4583-4613, 2020.

[22] Z. Xing, "An improved emperor penguin optimization based multilevel thresholding for color image segmentation," *Knowledge-Based Systems*, vol. 194, 105570, 2020.

[23] S. Meyyappan, S. Sathishbabu, N. Vinoth, M. Vijayakarthick, A.G. Ram, "Thresholding of Skin Melanoma Images based on Kapur's Entropy with Harmony Search Algorithm," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 11, pp. 716-726, 2020.

[24] S. Borjigin and P.K. Sahoo, "Color image segmentation based on multi-level Tsallis–Havrda–Charvát entropy and 2D histogram using PSO algorithms," *Pattern Recognition*, vol. 92, pp. 107-118, 2019.

[25] S. Kadry and V. Rajinikanth, "Grey Scale Image Multi-Thresholding Using Moth-Flame Algorithm and Tsallis Entropy," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 6, no. 2, pp. 79-89, 2020.

[26] M. Abd Elaziz, N. Nabil, R. Moghdani, A.A. Ewees, E. Cuevas, S. Lu, "Multilevel thresholding image segmentation based on improved volleyball premier league algorithm using whale optimization algorithm," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 12435-12468, 2021.

[27] M. Abd Elaziz, A.A. Heidari, H. Fujita, H. Moayedi, "A competitive chain-based Harris Hawks Optimizer for global optimization and multi-level image thresholding problems," *Applied Soft Computing*, vol. 95, 106347, 2020.

[28] J.N. Kapur, P.K. Sahoo, A.K.C Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Comput Vision Graph Image Process*, vol. 29, pp. 273–285, 1985.

[29] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, pp. 479–487, 1988.

[30] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," *Knowledge-Based Systems,* vol. 89, pp. 228-249, 2016, https://doi.org/10.1016/j.knosys.2015.07.006.

[31] S. J. Nanda, "Multi-objective moth flame optimization," In 2016 International conference on Advances in computing, communications and informatics (ICACCI), IEEE, 2016, pp. 2470-2476.

[32] M. Shehab, L. Abualigah, H. Al Hamad, H. Alabool, M. Alshinwan, A. M. Khasawneh, "Moth–flame optimization algorithm: variants and applications," *Neural Computing and Applications,* vol. 32, pp.9859-9884, 2020, https://doi.org/10.1007/s00521-019-04570-6.

[33] S.H.H. Mehne and S. Mirjalili, "Moth-Flame Optimization Algorithm: Theory, Literature Review, and Application in Optimal Nonlinear Feedback Control Design," *Nature-Inspired Optimizers*, vol. 811, pp. 143-166, 2020, https://doi.org/10.1007/978-3-030-12127-3_9.

[34] S. Grgic, M. Grgic, M. Mrak, "Reliability of objective picture quality measures," *Journal of Electrical Engineering*, vol. 55, no. 1–2, pp. 3–10, 2004.

[35] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600– 612.

[36] A. Hemeida, R. Mansour, M.E. Hussein, "Multilevel Thresholding for Image Segmentation Using an Improved Electromagnetism Optimization Algorithm," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 4, pp. 102-112, http://doi.org/10.9781/ijimai.2018.09.001

[37] S. Kadry, V. Rajinikanth, J. Koo, B.G. Kang, "Image multi-level-thresholding with Mayfly optimization," *International Journal of Electrical & Computer Engineering*, vol. 11, no. 6, pp. 5420-5429, 2021.

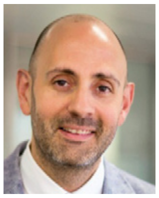
[38] V. Rajinikanth, S.M. Aslam, S. Kadry, O. Thinnukool, "Semi/Fully-Automated Segmentation of Gastric-Polyp Using Aquila-Optimization-Algorithm Enhanced Images," Cmc-Computers Materials & Continua, vol. 70, no. 2, pp. 4087-4105, 2022.

V. Rajinikanth

V. Rajinikanth is a Professor in Department of Electronics and Instrumentation Engineering at St. Joseph's College of Engineering, Chennai 600119, Tamilnadu, India. He has published more than 125 papers and authored/edited 7 books in the field of medical data assessment. His main research interests include Heuristic algorithm based optimization, Image thresholding, Machine learning and Deep learning.

Seifedine Kadry

Seifedine Kadry (Senior Member, IEEE) received the bachelor's degree from Lebanese University, in 1999, the dual M.S. degree from Reims University, France, and EPFL, Lausanne, in 2002, the Ph.D. degree from Blaise Pascal University, France, in 2007, and the H.D.R. degree from Rouen University, in 2017. He is an IET Fellow, IETE Fellow. His research interests include data science, education using technology, system prognostics, stochastic systems, and applied mathematics. Currently, he is a full professor of data science at Noroff University College, Kristiansand, Norway.



Rubén González Crespo

Dr. Rubén González Crespo has a PhD in Computer Science Engineering. Currently he is Vice Chancellor of Academic Affairs and Faculty from UNIR and Global Director of Engineering Schools from PROEDUCA Group. He is advisory board member for the Ministry of Education at Colombia and evaluator from the National Agency for Quality Evaluation and Accreditation of Spain (ANECA). He is member from different committees at ISO Organization. Finally he has published more than 200 paper in indexed journals and congresses.

Elena Verdú

Elena Verdú received the master's and Ph.D. degrees in Telecommunications Engineering from the University of Valladolid, Spain, in 1999 and 2010, respectively. She is currently an Associate Professor with the Universidad Internacional de La Rioja (UNIR), where she is also a member of the Research Group "Data Driven Science." For more than 15 years, she has worked on research projects at both national and European levels. Her research has focused on e-learning technologies, intelligent tutoring systems, competitive learning systems, data mining, machine learning, natural language processing, image and speech processing, and expert systems.

Mapping and Deep Analysis of Image Dehazing: Coherent Taxonomy, Datasets, Open Challenges, Motivations, and Recommendations

Karrar Hameed Abdulkareem[1,2]*, Nureize Arbaiy[2], Zainab Hussein Arif[3], Mohammed Nasser Al-Mhiqani[4], Mazin Abed Mohammed[5], Seifedine Kadry[6], Zaid Abdi Alkareem Alyasseri[7]

[1] College of Agriculture, Al-Muthanna University (Iraq)
[2] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn (Malaysia)
[3] College of Computer Science and Information Technology, University of Al-Qadisiyah (Iraq)
[4] Center for Advanced Computing Technology, Faculty of Information Communication Technology, Universiti Teknikal Malaysia Melaka (Malaysia)
[5] College of Computer Science and Information Technology, University of Anbar, 11, Ramadi, Anbar (Iraq)
[6] Department of Applied Data Science, Norrof University College, 4608 Kristiansand (Norway)
[7] ECE Department-Faculty of Engineering, University of Kufa, P.O. Box 21, Najaf (Iraq)

Received 5 October 2021 | Accepted 24 November 2021 | Published 29 November 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

Our study aims to review and analyze the most relevant studies in the image dehazing field. Many aspects have been deemed necessary to provide a broad understanding of various studies that have been examined through surveying the existing literature. These aspects are as follows: datasets that have been used in the literature, challenges that other researchers have faced, motivations, and recommendations for diminishing the obstacles in the reported literature. A systematic protocol is employed to search all relevant articles on image dehazing, with variations in keywords, in addition to searching for evaluation and benchmark studies. The search process is established on three online databases, namely, IEEE Xplore, Web of Science (WOS), and ScienceDirect (SD), from 2008 to 2021. These indices are selected because they are sufficient in terms of coverage. Along with definition of the inclusion and exclusion criteria, we include 152 articles to the final set. A total of 55 out of 152 articles focused on various studies that conducted image dehazing, and 13 out 152 studies covered most of the review papers based on scenarios and general overviews. Finally, most of the included articles centered on the development of image dehazing algorithms based on real-time scenario (84/152) articles. Image dehazing removes unwanted visual effects and is often considered an image enhancement technique, which requires a fully automated algorithm to work under real-time outdoor applications, a reliable evaluation method, and datasets based on different weather conditions. Many relevant studies have been conducted to meet these critical requirements. We conducted objective image quality assessment experimental comparison of various image dehazing algorithms. In conclusions unlike other review papers, our study distinctly reflects different observations on image dehazing areas. We believe that the result of this study can serve as a useful guideline for practitioners who are looking for a comprehensive view on image dehazing.

Keywords

Image Dehazing, Image Defogging, Image Dehazing Algorithms, Evaluation, Image Quality Assessment, Dataset.

DOI: 10.9781/ijimai.2021.11.009

I. Introduction

Computer vision is an interdisciplinary research [1] that is relevant to a wide range of applications that can influence our daily life, such as vehicle navigation, surveillance, and traffic monitoring [2]. Although computer vision applications are popular in indoor environments, they remain constrained in outdoor environments [3]. The degradation of outdoor scene images could be attributed to various reasons, but the main reason is turbid weather. Bad weather conditions could be dynamic (rain and snow) or steady (fog, mist, and haze) depending on the kinds and sizes of particles in the atmosphere and their density in the air [3]. The images capture haze weather are usually degraded in terms of fidelity and low contrast because light is scattered and absorbed as it travels in bad weather conditions. Consequently, most outdoor applications that rely heavily on the quality of input images do not work efficiently because of degraded images [4]. Thus, the enhancement of image quality in bad weather conditions is critical in countless computer vision applications [1].

* Corresponding author.

E-mail address: khak9784@mu.edu.iq

Haze is an atmospheric effect that sets a gray color over a scene, thereby decreasing the visibility in outdoor scene images [5]. Haze is also considered one of the main causes of accidents in different environmental mediums, such underwater, air, and land [6]. Particles such as smoke and moisture, which usually spread in the air, scatter the light that propagates through the atmosphere and cause the formation of haze [7]. The process of eliminating haze effects from outdoor images and restoring fidelity details is called dehazing and it is often considered an image enhancement technique [5]. However, it is unlike traditional contrast enhancement methods because the degradation of image pixels induced by the presence of haze depends on the distance between the object and the acquisition device and the regional density of the haze [8].

Fog formation has two aspects, namely, attenuation and airlight. Attenuation reduces contrast, whereas airlight increases whiteness in a scene. In attenuation, the light rays that propagate from a specific scene point due to the scattering of atmospheric particles are attenuated [9]. The light propagating from the source is scattered on its way to the camera and inserts whiteness in the scene or causes color distortion, that is, airlight [10], [11]. Furthermore, the variations of the effects of airlight and attenuation are restricted to the distance between the scene point and the device (e.g., camera). Therefore, the accuracy of the restoration of degraded images mainly depends on the remap concept, that is, the estimation of the depth or airlight map [9].

An image defogging algorithm must be designed to improve the environmental adaptability of visual systems. Many improved defogging algorithms based on physical models have been proposed for use in outdoor scenes [6]. Some video and image defogging algorithms have also been proposed for real-world traffic surveillance scenes [12]. Most existing defogging algorithms are aimed at removing fog from land images. However, few studies on sea and air images exist. In some works, the image defogging algorithm was simply divided into two categories according to whether a physical model was used or not [13]. The first category is image restoration based on a physical model [14], [15], and the other is based on image enhancement [16], [17]. The image restoration method establishes a physical imaging model on the basis of the cause of image degradation under foggy conditions. Under this category, the algorithms must estimate the parameters of the physical model, such as the atmospheric light and transmission (depth) [18]. An image can be restored by inversely solving the physical model. Image restoration algorithms are aimed at obtaining a natural and clear image with good visibility while maintaining good performance in terms of color restoration. The second category of defogging algorithms is based on image enhancement and does not consider the physical imaging model of foggy conditions. Algorithms under this category attempt to use various image enhancement methods to enhance the contrast and visibility of foggy images [6]. In recent years, fusion-based defogging algorithms that enhance images by fusing multiple input images have received considerable attention [19], [20]. Thus, fusion-based defogging algorithms can be regarded as the third category of defogging algorithms. However, image restoration algorithms based on physical models can be divided into two categories according to the number of images used: image restoration based on multiple images [3] and image restoration based on a single image [10], [21] .

To prove the efficiency of a particular algorithm, evaluation and benchmarking are necessary steps in image dehazing. Image quality assessment methods enable us to compare the performance of different image dehazing algorithms. Various foggy scenes have been made available to test the usefulness of image dehazing algorithms [22], [23]. Most forms of assessment are equivalent on several foggy scenes [6], [8], [24], [25]. For example, in [6] the authors considered a variety of evaluation scenes, including inhomogeneous, homogeneous, and dark foggy scenes to test the efficiency of algorithms. Therefore, the advantages and demerits of each algorithm should be considered within each context. Under different hazy scenes, several algorithms can work properly, such as those proposed in [31], [33]. Therefore, comparing these algorithms from only one perspective is unfair [34]. The efficiency of image dehazing algorithms also needs to be evaluated by using trustworthy approaches [24], [37]. In this case, how several algorithms can be evaluated and how the best algorithm is selected through an effective approach warrant further investigation. Different image quality assessment methods have been proposed for evaluation and benchmarking of image dehazing algorithms. So far, there are no reliable means to measure the quality of the image dehazing algorithms [24], [37].

Our study attempts to highlight several aspects within the image dehazing area, and the study contributions can be summarized as follows:

- We highlight the developments in real-time image dehazing algorithms.
- We sum up significant achievements by other researchers in response to image dehazing needs.
- We draw attention to evaluation methodologies and datasets.
- A comprehensive evaluation of experiments is presented based on different algorithms as well as different foggy scenes.
- We propose a taxonomy that maps the existing literature in a well-ordered body and defines various research lines in the image dehazing field. We believe that the outcomes are beneficial to other researchers.

The presents study has been organized into different sections. Section II introduces the details of the systematic review procedure. Section III provides results of the adopted systematic review protocol. Section IV focuses on technical aspects where different reviewed works have been implemented and evaluated based on well-known metrics. Section V discusses with details all achieved results from the proposed taxonomy as well as the evaluation experiments. Section VI highlights the constraints of the present review study. Section VII concludes on the contributions of this study and maps the addressed challenges with achieved outcomes.

II. Systematic Review Protocol

A. Information Sources

In terms of systematic search, we selected three of the most popular online search engine databases: Web of Science (WOS), ScienceDirect (SD), and IEEE Xplore Digital Library. The selection was established according to the index that eases and formulates a simple and complex search query and especially monitors numerous journals and conference papers in the sciences, including computer science and social science. This selection was aimed at including as much literature as possible that covers the maximum number of articles related to image dehazing and technical ones. It was also aimed at providing a holistic view of researchers' achievements in a broad but pertinent variety of disciplines.

B. Study Selection

The study selection technique implied an exhaustive search of related articles involving two steps. First, irrelevant and duplicated articles were excluded by means of scanning the titles and abstract. Second, the articles scanned in the previous step were filtered through full text reading. The same eligibility criteria were applied to the two stages.

C. Search

The article search process was launched on 08 March 2018, and the search query was used on the IEEE, WOS, and SD databases via their search boxes. In all the mentioned databases, searching was carried out using keywords related to terminologies ("image dehazing" OR "image defogging" OR "image dehaze" OR "image defog" OR "hazy image" OR "foggy image" OR "video dehazing" OR "video defogging" OR "haze removal" OR "fog removal") that were combined later through the "AND" operator with the following keywords ("Evaluation" OR "Benchmarking" OR "Assessment" OR "Measurement"), as shown in Fig. 1. Advanced search preferences in each engine were utilized to exclude the chapters of books and other types of documents and to include only the relevant journals and conference papers. Furthermore, we considered the studies that were most undoubtedly immersed in the latest and suitable scientific research related to our study.

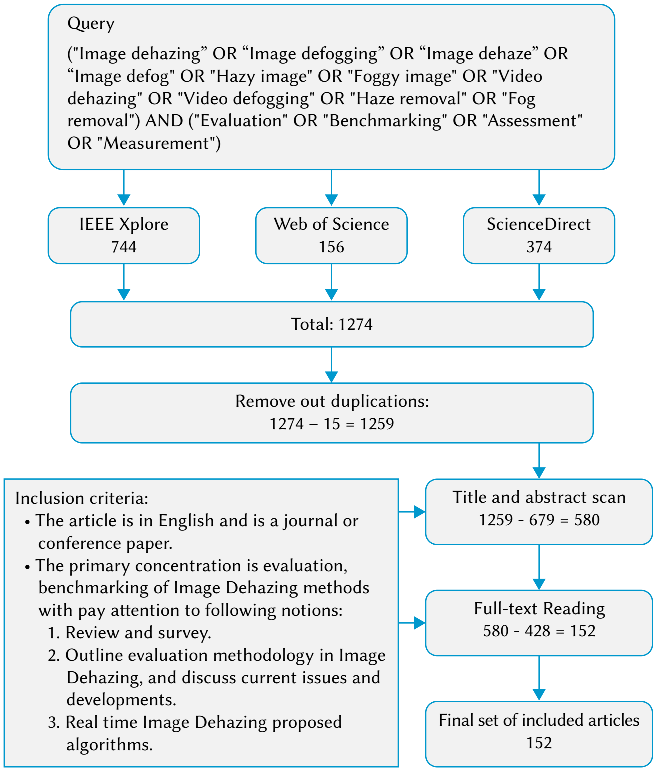

Query
("Image dehazing" OR "Image defogging" OR "Image dehaze" OR "Image defog" OR "Hazy image" OR "Foggy image" OR "Video dehazing" OR "Video defogging" OR "Haze removal" OR "Fog removal") AND ("Evaluation" OR "Benchmarking" OR "Assessment" OR "Measurement")
IEEE Xplore 744
Web of Science 156
ScienceDirect 374
Total: 1274
Remove out duplications: 1274 – 15 = 1259
Inclusion criteria:
• The article is in English and is a journal or conference paper.
• The primary concentration is evaluation, benchmarking of Image Dehazing methods with pay attention to following notions:
1. Review and survey.
2. Outline evaluation methodology in Image Dehazing, and discuss current issues and developments.
3. Real time Image Dehazing proposed algorithms.
Title and abstract scan 1259 - 679 = 580
Full-text Reading 580 - 428 = 152
Final set of included articles 152

Fig.1. Study selection diagram.

D. Eligibility Criteria

Entire articles that match the criteria shown in Fig. 1 were included. We set the primary goal as mapping the compass of research on image dehazing into a wide-range and coarse-grained taxonomy of three groups. The groups were procured from a comprehensive pre-review of the existing literature with no restriction. To eliminate replicated articles, we excluded the articles that failed to match the eligibility criteria. The exclusion criteria covered the following consecutive points: (1) the article is non-English; (2) the focus is on limited aspects of image dehazing, such as non-real-time methods.

E. Data Collection Process

The well-known EXCEL® software was employed to coordinate the final set of articles that was assembled through the study selection with the corresponding initial categories. We achieved multiple full text readings of the articles included to underline the importance of details and comments on the revised studies and in a running classification of articles in a polished taxonomy. The highlighted details and comments were found in the body of the texts (corresponding to the authors' desired style). The significant outcomes were summarized, tabulated, and described. Word and Excel forms were used to save information, such as article lists, relevant online source databases, summary and description tables, study types, review sources, utilized datasets, dataset types, evaluation types, evaluation metrics, and different related figures. These details were presented in a manner in which the auxiliary materials could serve as a full reference for the results. They are defined in the next section.

III. Results

The first run of the search query filtered 1274 articles with the following details: 744 articles from IEEE Xplore search engine, 374 articles from SD, and 156 articles from WOS over a period of 13 years (2008–2021). Fifteen articles were duplicates. Through title and abstract scanning, 679 articles were excluded as non-related ones, resulting in 580 articles. After the full text reading step, 428 articles were excluded. Finally, 152 articles were included in the final set of articles. These articles were examined carefully to obtain a generic research overview of this emerging area. Nevertheless, a variety of studies have focused on the same area. The articles were categorized on the basis of the aim of the study and utilized to serve the process of taxonomy formation. Fig. 2 shows the proposed taxonomy for reviewing the research articles that focused on image dehazing. Consequently, three types of article categories were identified in the obtained taxonomy. First, out of 152 articles, only 55 of them focused on various studies on image dehazing, such as the comparative study of different image dehazing algorithms, multiple evaluation methods and proposed metrics, and different datasets based on diverse scenes and circumstances. Second, 13 studies conducted a review and survey, reviewing different aspects such as multiple methods based on dark channel models, underwater image dehazing metrics and methods, suitable methods for driver assistance systems, and comprehensive investigations into the image dehazing field. Third, 84 articles were focused on the development of methods for real-time scenarios.

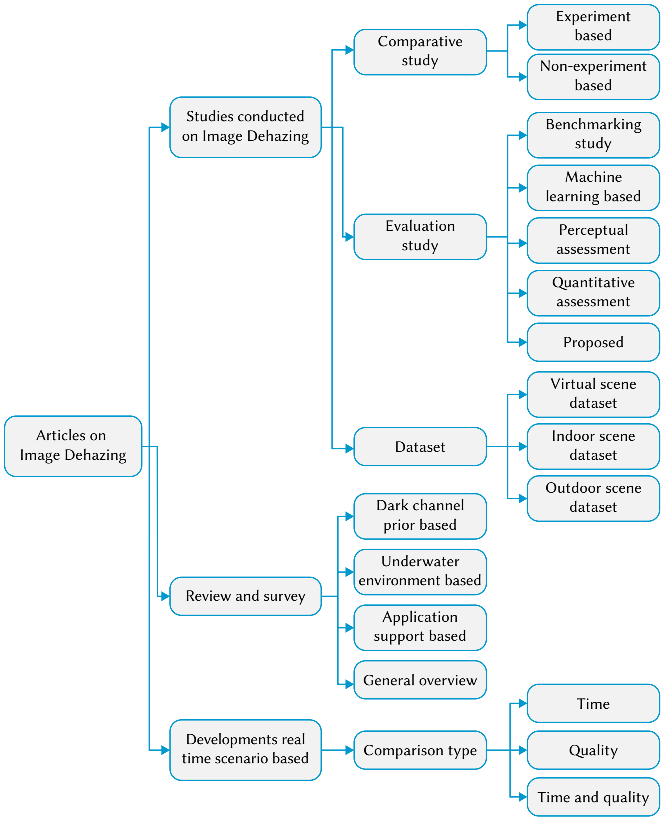


Fig. 2. Taxonomy of research literature on image dehazing.

Through outlines, we indicated the generic categories of articles and revised the classification into a literature taxonomy, as shown in Fig. 2. In this figure, we also illustrate numerous subcategories of the main classes without any overlap. The following sections describe the perceived categories and provide simplified associated statistics.

1. Studies Conducted on Image Dehazing

Generally, the image dehazing literature focuses mainly on the development of new methods. The second largest article group comprises diverse studies (55/152) on image dehazing.

We divided the works included into three main subcategories: comparative study (3/55), evaluation study (18/55), and dataset-based study (34/55). Comparative studies are experiment-based [13], [26] and non-experiment based studies [27]. Three studies compared the prevalent approaches in this area through the implementation of methods and using the most common parameters for critical analysis. One study compared several well-known visibility enhancement techniques without further implementation.

More articles are found in the evaluation category (18/55), which is basically introduced through different evaluation types and techniques. A new direction in the image dehazing domain has been presented in the recent years by authors in [28]-[30], where image dehazing algorithms are evaluated and selected based on principle of multi-criteria decision making (MCDM). An evaluation study [31] compared the performance of different techniques for underwater autonomous vehicles, with the preferred characteristics being the reduced need for additional hardware, short computation time, and simple inputs. Machine learning techniques for assessing the quality of dehazed images were introduced in two studies that respectively presented a novel quality assessment framework for the performance ranking of image dehazing algorithms [32] and a relative quality ranking between enhanced images instead of absolute quality scoring for a single enhanced image [33], [34]. Existing works reported that image quality assessment methods are mainly divided into two types. The first type is subjective (perceptual) assessment involving psychophysical experiments in which human observers are asked to grade a set of images according to a given quality criterion to offer agreement among observers on the quality of haze-free and dehazed images [35], [36]. The other type of well-known image quality assessment is quantitative (objective) assessment, which is the essential element in the evaluation of perceived image quality [37]. Quantitative evaluation is more structured and reliable than subjective evaluation. This tool usually uses the objective criterion and offers an identified procedure for measuring image quality. Moreover, this method deals only with numeric results, and any user who wants to use this method will obtain the same result [22]. Even with the popularity of quantitative assessment, other consistent methods of objective evaluation must be developed to provide accurate judgements on image dehazing algorithms [24]. Finally, several methods and metrics of image quality assessment have been proposed. First, a method was developed on the basis of the circularly symmetric Gaussian normalization procedure's visible edge feature, which does not require exposure to distorted images priori and training [23]. Second, a quantitative assessment method based on two optimization objectives was introduced with consideration of several aspects for evaluation, such as the effects of color distortion of the dehazing process and halo artifacts in restored images [38]. Third, three new methods (contrast measurement index (e), image naturalness index (CNI), and colorfulness index (CCI)) were combined to assess the defogging algorithm [39]: a new metric based on underwater scattering and absorption aspects [40], an evaluation metric combining contrast degree with structural similarity [41], and a novel no-reference haze assessment method based on haze distribution for remote sensing images [42].

In terms of support for the evaluation process and development of new image dehazing algorithms, the largest group of articles have been found in the datasets category (34/55) which are presented in three forms. The virtual scene dataset was basically created by utilizing computer graphics to produce an enormous number of hazy images (2000 images) based on road scenes with different levels of fog [43]. Indoor scene datasets were established through real scenes inside a room with a fog machine to generate 9 images [44], 1400+ images [45],and controlled underwater environment images using milk to obtain the turbidity in a water tank [46]. The outdoor scene datasets were designed with two scenarios, namely, a database that consists of natural scenes in uncontrolled outdoor conditions (5640 images [47] and 3464 images [48]) and a synthetic outdoor dataset created through synthesized haze in real images with complex and multiple scenes [49].

To enrich the development of image dehazing methods and the practice of image quality assessment, we reveal several types of datasets in this study. The variations of datasets depend on scene conditions and environmental domains. Scene types can be classified according to circumstances, such as indoor, outdoor, and road traffic scenes. The haze removal process requires two types of images, namely, hazy images for removing the noise and haze-free images for measuring the volume of enhancement. Thus, providing images reflecting various weather and illumination conditions, such hazy weather, foggy weather, poor illumination, and normal daylight, is a vital factor in the image dehazing practice. On the one hand, because atmospheric light varies between over-land and underwater scenes, some datasets are built on the basis of this context; examples include datasets of real underwater scenes [31] and synthesized datasets of underwater images taken in a water tank [40]. On the other hand, datasets have been classified according to whether they were built on real or virtual scenes. For real scenes, most images are taken using a camera based on indoor or outdoor natural images [47], [48], and these real scenes could be utilized for synthesizing new ones through different equipment for generating haze [45], [50]. For virtual scene-based datasets, images are usually generated using computer graphic techniques to render scenes [43], [51].

Further details on image dehazing datasets are presented in Table I. Our study provides several details about existing datasets, such as a reference using a dataset, total number of images, and sources and types of datasets involved. Although realizing different aspects of datasets is significant in image dehazing, multiple algorithms must be evaluated, and a new image quality assessment methodology must be proposed because authors are required to verify the efficiency of the developed methods in terms of enhancing and restoring images. Furthermore, the main goal of developers and researchers is to provide a public dataset that can be used for dedicated purposes, such as in validating and evaluating their methods.

Noticeably, image dehazing researchers are divided into two groups: those who built their own datasets and those who used public datasets or datasets from specific studies. In general, most researchers prefer natural outdoor scenes. Others favor the use of more datasets in their studies, specifically virtual and real image datasets. In terms of datasets based on a specific environmental domain, most existing datasets are on over-land scenes, and few are based on underwater scenes. However, most studies have widely used the FIRDA dataset because it involves different aspects, such as various kinds of foggy scenes (uniform, heterogeneous, cloudy, and cloudy heterogeneous fog), which can enrich the evaluation scenario from multiple perspectives. The dataset also presents a full reference scenario (clear and foggy images), which is the most desired aspect because achieving it is difficult in real world scenes and recording such images is not feasible due to the variations of illumination conditions [45].

TABLE I Dataset Statistics

Ref	Dataset	Over-land	Over-water	Underwater	Real	Synthesis	Indoor	Outdoor	Source
[31]	Dataset1 = 19 images (Rocks) Dataset2 = 94 images (sand and Rocks) Dataset3 = 100 images (shallow corals) Dataset4 = 99 images (medium corals) Dataset5 = 100images (deep corals)	✗	✗	✓	✓	✗	✗	✗	[52]
[53, 54]	FRIDA dataset = 90 images FRIDA2 dataset = 330 images FRIDA3 dataset = 264 images (publicly available)	✓	✗	✗	✗	✓	✗	✓	http://perso.lcpc.fr/tarel.jean-philippe/bdd/frida.html
[6]	WILD (Weather and Illumination Database) dataset = 3000 images (publicly available)	✓	✗	✗	✓	✗	✗	✓	http://www.cs.columbia.edu/CAVE/software/wild/index.php
[55], [43], [56], [57]	Dataset (Fattal, 2014) 11 haze images (publicly available)	✓	✗	✗	✓	✗	✓	✓	http://www.cs.huji.ac.il/~raananf/projects/dehaze_cl/results/index_comp.html
Frequency	-	79 %	3 %	17%	47%	50%	29%	61%	-

By contrast, the real image dataset of Fattal shares almost the same importance as the FIRDA dataset. In general, real images are more valid for real scenarios than synthetic datasets [45].

According to Table I and Appendix A, the majority of the existing datasets belong to over-land scenes where 79 % of datasets are constructed based on this type. Due to complexity of the environment and procedure to collect the data in underwater environment only 17% of datasets are founded belong to underwater scene. However, only one study was found that belongs to new direction in the construction of image dehazing datasets which is over-water scene. Regarding the evaluation experiment based on real and synthesis images; it was observed that only 3% are the differences where more articles were found to belong to synthesis type. Also, images based on outdoor scenes are much more than indoor images, where 61% of reported datasets belong to outdoor type.

2. Review and Survey

The review articles on image dehazing aimed to highlight new developments and provide a comprehensive view for image dehazing followers. The smallest article group in the taxonomy comprises the reviews and survey group of the literature (i.e., 13 out of 152 articles). These articles were classified on the basis of what the algorithms support, such as application support. Similar to this context, dark channel prior (DCP) is the most popular image dehazing model because of its adequate performance and potential for improvements and applications; the authors in [58] studied approaches on the basis of the DCP model. Three articles [59]-[61] reviewed the latest methods that have been effectively applied in the underwater environment, achieved good underwater image dehazing and color restoration performance with different methods, developed an underwater image color evaluation metric, and highlighted different underwater image applications. To find a suitable approach for vision-based driver assistance systems, an article [62] in the existing literature reviewed state-of-the-art image enhancement and restoration methods.

Most survey and review articles are based on the general view of image dehazing (7/14). These articles examined and summarized different methods of image dehazing, such as image enhancement methods, physical model restoration methods, and fusion-based visibility enhancement techniques [1], [2], [8], [63]. These methods were also categorized on the basis of the type of technique used to acquire information required by the image restoration process; examples include multiple image methods, polarizing filter-based methods, methods with known depth, and single-image methods [55], [64]. Finally, the authors in [6] reviewed the detection and classification method of foggy images and summarized the objective image quality assessment methods that have been widely used to compare different defogging algorithms.

Further analysis is presented in Table II which shows that most of the review articles on image dehazing were classified as other existing studies based on certain concepts. Several articles classified image dehazing algorithms into the following three forms on the basis of input type required by the dehazing process: single input image, multiple images, and additional information approaches [58], [55]. Fog, haze, smoke, mist, rain, and dust are weather conditions provided by a certain dataset, and according to these conditions, several datasets were classified [1], [2]. Most review studies focused

TABLE II. Critical Analysis of Review Studies on Image Dehazing

Ref	Input type	Dataset classification	Quality assessment	Application classification	Metric classification
[1]	✗	✓	✗	✗	✗
[2]	✗	✓	✓	✗	✗
[6]	✓	✗	✓	✗	✓
[8]	✓	✗	✓	✗	✓
[58]	✓	✗	✗	✗	✗
[59]	✗	✗	✗	✗	✗
[60]	✓	✗	✓	✓	✗
[62]	✗	✗	✗	✗	✗
[63]	✗	✗	✓	✓	✗
[64]	✓	✗	✓	✗	✗
[55]	✓	✗	✓	✗	✗
[61]	✓	✓	✓	✗	✓
[65]	✗	✗	✓	✗	✗
Frequency	**53%**	**23%**	**69%**	**15%**	**23%**

TABLE III. Critical Analysis of Real-time Image Dehazing Algorithms

	Approach					Evaluation		Data type				
Ref	Image restoration	Image enhancement	Image fusion	Hybrid	Technique	Subjective	Objective	Image	video	Image and video	Scene type	Application support
[5]	✓	×	×	×	Gaussian surround filter and DCP	✓	✓	✓	×	×	General	Not specified
[9]	×	×	×	✓	Anisotropic diffusion	×	✓	✓	×	×	General	Not specified
[86]	✓	×	×	×	DCP and gray projection	×	✓	×	×	✓	General	Not specified
[107]	✓	×	×	×	Machine learning	✓	✓	✓	×	×	General	Not specified
[108]	✓	×	×	×	Machine learning	×	✓	✓	×	×	General	Not specified
Frequency	70%	10%	2%	16%	-	23%	95%	80%	5%	13%	-	-

on classifying approaches and methods for image dehazing, whereas several studies presented an image quality assessment of certain haze removal algorithms [55], [64]. Many applications have taken advantage of employing image dehazing algorithms as a preprocessing step, but only a few studies have classified the applications related to this area [60], [63]. Finally, in terms of criteria for evaluation of certain image dehazing algorithms, numerous metrics have been reported in the existing literature, but few studies have classified these metrics according to the most critical evaluation criteria [6], [8].

However, input type is an important aspect of image dehazing in terms of defining the complexity of steps and the type of procedure to define the transmission map or estimate the airlight. Table II shows that many studies (53%) have considered input type as a classification aspect. Furthermore, few studies (23%) have considered types of weather conditions that should provide through experiment of data acquisition, thus presenting more complex scene is very beneficial in efficiently verifying the performance of multiple image dehazing algorithms. In this direction, our study presents many types of datasets and more information about these datasets (see Table I), enabling other researchers to select the most appropriate one for a particular study. Moreover, quality assessment is a vital step in evaluating the performance of certain algorithms, and it facilitates the selection of the best algorithms for specific scenarios. In this direction, many studies (69%) have highlighted the types of evaluation approaches and discussed details of the evaluation. In addition, many applications have been based on the image dehazing concept, but only a few studies (15%) have classified these applications. In addition, providing an umbrella for the types of metrics that can be used for evaluation scenarios is significant; it can also define the most suitable metric for a specific case study. However, only few (23%) studies considered this matter.

3. Development of Methods Based on Real-time Scenario

Apparently, most research works on image dehazing are development articles (84/152) dedicated to improving the process of dehazing through the enhancement of the quality of degraded images and the increase of the speed of restoration. Typically, proposing a new method requires an evaluation process to measure the effectiveness and efficiency of the proposed approach. Thus, in the current work, new algorithms of image dehazing were compared with other algorithms in terms of execution time and quality. We classified development articles according to comparison details provided through the literature, especially in the quantitative evaluation section. A total of 11 out of 84 articles mentioned time as the sole evaluation criterion; it is the most preferred indicator in real-time scenarios [66], [67]. A total of 10 out of 84 articles stated quality as the performance comparison metric [51],[68]-[70]. Finally, most comparison settings are based on time and quality criteria, that is, 60 out of 69 articles. As shown in Table III, articles were classified into many aspects. First, numerous algorithms support certain types of applications, such as driver assistance systems [71]-[73], road sign detection [74]-[76], monitoring of power plants [77], optical systems [78], surveillance applications [79], [80], embedded systems [81], unmanned aerial vehicles [82], and car vision [83]. Second, some algorithms concentrate on hardware implementation or utilize a specific hardware architecture, such as heterogeneous multi-cores [71], field-programmable gate arrays (FPGAs) [81], [84], and a seven-stage pipeline hardware architecture [85]. Third, through experiments, several types of data were examined, and they include image and video [86]-[89], video sequence only [90], or image only [91], [92]. Fourth, as mentioned, the evaluation of a certain algorithm was divided into two types, namely, subjective and objective; most algorithms were objectively evaluated [93], [94], and only a few studies adopted a subjective approach on the basis of user observations that rate the perceived quality of tested images [5], [87], [95], [96]. Fifth, image dehazing algorithms were proposed using different approaches and techniques. These approaches could be based on a physical model [9], [93], a non-physical model [97], image fusion [98], [99], and approaches that combine image enhancement and restoration [100], [101] or image restoration and fusion [102], [103]. Finally, because of the satisfactory performance of the DCP, it has been adopted in many image dehazing algorithms [89], [104]. Moreover, only a few algorithms have been based for other techniques, such as machine learning [105], [106].

According to the **Table III and Appendix B**, due the advantages of depth estimation for image dehazing physical model most of the studies (70%) are conducted based on restoration approach. Minimal studies have adopted other approaches such as image enhancement (10%) and fusion (2%).

However, a new trend is presented by few studies (16%) that used the image restoration approach relative to image enhancement or image fusion. In some cases, these studies leveraged the image enhancement procedure as a post processing step with image restoration or image fusion. Furthermore, because of its simplicity and speed, DCP (model) has been widely employed in image dehazing algorithms. In terms of evaluation, most researchers only (23%) try to avoid the subjective method, which involves user opinion, because of its disadvantages. They tend to prefer to deal with the structured method, which involves specific criteria (objective method) where almost 95% of articles are include a quantitative evaluation approach. In terms of data tested with the algorithm, images have been widely used (80%) because they require less processing than videos (many frames) do. Only a few studies involved special hardware implementation, such as FGPA, to provide full real-time scenarios that are based on embedded systems. Finally, as mentioned in the Motivations section, the image dehazing principle has been widely adopted in various applications. However, the types of application supported by many algorithms are not specified, thus contributing to the difficulty of selecting a suitable algorithm for certain applications. To mention, most of the existing studies are preferred to use general hazy scenes in other word more than specific hazy images such as inhomogeneous, homogenous, dark, and sky in order test the validity of certain proposed algorithm. On other hand, some algorithms are dedicated for enhancement of specific hazy image such as sky or inhomogeneous or daytime rather than night-time. The most surprising part is that the principle of image dehazing is used in different case studies that not involved real haze characteristics such as TV industry, Biometric, Steganography, and nondestructive testing (NDT). Meanwhile, several algorithms support driver assistance systems, agriculture monitoring, railway industry, mobile cloud of smart city, and so on. Therefore, existing algorithms need more experiments on video datasets to validate their performance in terms of frame sequence processing and on more embedded systems to verify their suitability for real-time applications. Similar to other researchers, we recommended the use of objective evaluation rather than subjective evaluation.

To highlight and understand the trends in the research literature, which is one of our study's contributions, Fig. 3 illustrates the number of publications gathered from the literature along with the corresponding search engine types and presents further content analysis. The statistics for the articles are covered in the final set (152). As shown in Fig. 3, significant attention was given to the development of new methods for image dehazing using real-time scenarios.

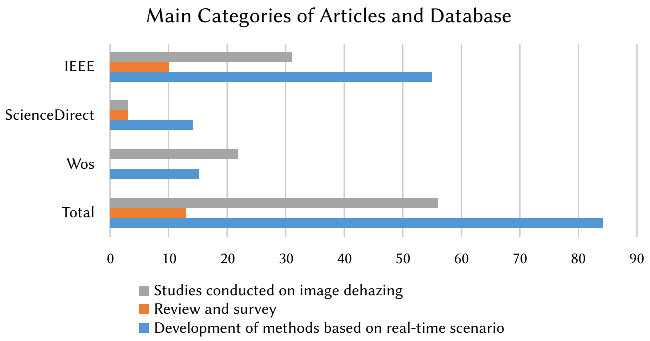


Fig. 3. Number of articles by their main categories and database sources.

Fig. 4 specifies the number of articles according to category and year of publication. Apparently, significant efforts have been exerted to explore image dehazing in recent years, particularly in development studies and review and survey articles. As mentioned previously, studies on image dehazing showed 55 papers, the review and survey category showed only 13 articles, and the category on the development of real-time scenario-based algorithms showed 84 articles.

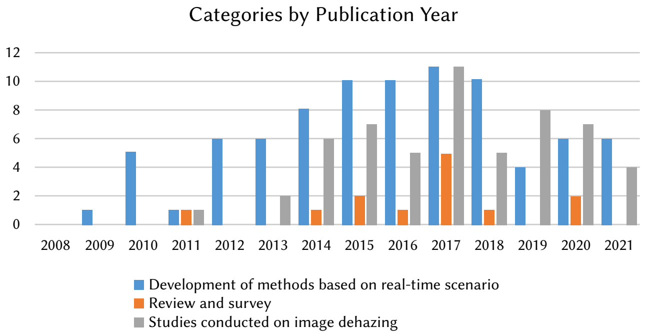


Fig. 4. Number of articles in each category by year of publication.

Fig. 5 shows the distribution of articles within each country. China was the country more focused on image dehazing with 74 contributions from different Chinese organizations and universities. This could be relevant to existence of bad weather during different seasons as well as the smoke or haze emission from factories. However less attention for image dehazing topic has been found by different countries such as Australia, Austria, Canada, Norway, and so on.

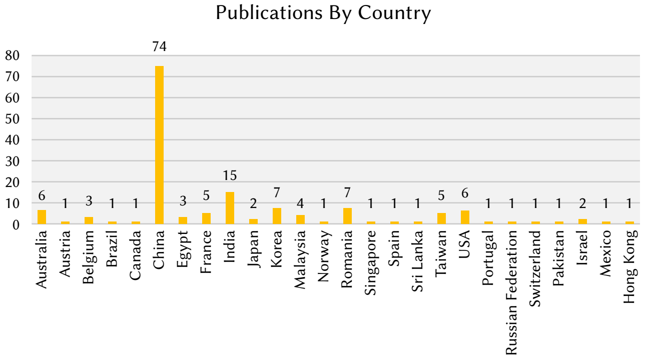


Fig. 5. Number of articles according to country.

IV. Evaluation of Experimental Results

In this section, some image dehazing algorithms were compared via image quality assessment experiments. Seventeen image dehazing algorithms are included in this experiment such as **Dehazenet** [68], **MSCNN** [140], **Colores** [141], **Zhu** [4], **Multi-band** [142], **CO-DHWT** [143], **Meng** [144], **Liu** [145], **Berman** [146], **BF** [147], **WBCID** [148], **GF** [149], **JBF** [150], **Kim** [184], **NHR** [151],**He et al.** [10], and **Tarel** [152].The evaluation experiment is conducted based on the two datasets LIVE Image Defogging Database [81] and RESIDE [66]. According to [6], [28], [29] the evaluation of image dehazing algorithms based on different hazy scene characteristics provides comprehensive image quality assessment. Thus, the potentials of a certain algorithm can be measured with different and more complex scenes. Along with this, four main evaluation scenes are included in our experiment namely inhomogeneous foggy scene, homogeneous foggy scene, dark foggy scene, and sky foggy scene (see Fig. 6 and Fig. 7). Also, the evaluation criteria are selected based on recommendation from other studies specifically [6], [24], [25]. These criteria are divided into quality and time. Where each algorithm will be measured based on exaction time and each of e, r, Σ, HCC, SSIM, and UQI. Further details about criteria can be founded in the three mentioned references.

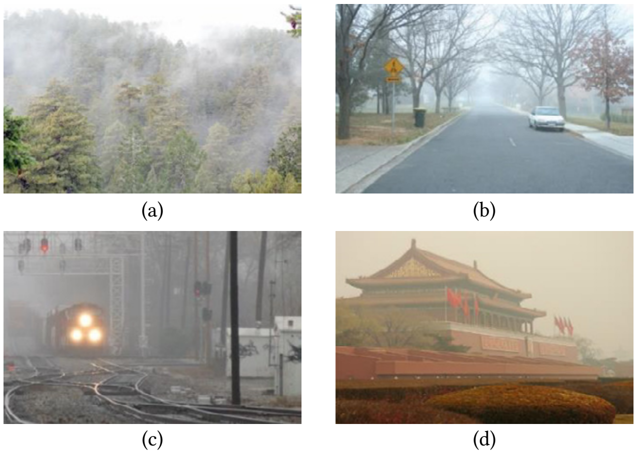
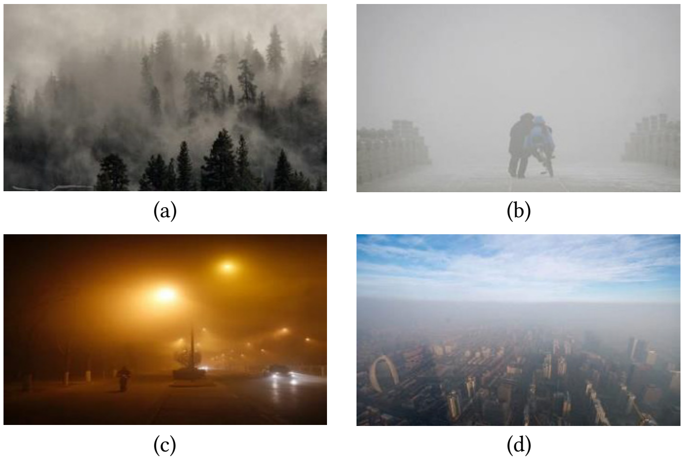

(a)
(b)
(c)
(d)

Fig. 6. LIVE Image Defogging Database: (a) inhomogeneous foggy scene, (b) homogenous foggy scene, (c) dark foggy scene, and (d) sky foggy scene.



(a) (b) (c) (d)

Fig. 7. RESIDE Database: (a) inhomogeneous foggy scene, (b) homogenous foggy scene, (c) dark foggy scene, and (d) sky foggy scene.

According to the Table IV, WBCID algorithm scored the best performance only in each of SSIM, UQI, and Time complexity. On other hand, Tarel have shown the best performance in terms e and Σ criteria. However, all other algorithms have scored better performance than WBCID and Tarel in other criteria.

However, **Appendix C**, NHR algorithm scored the best performance only in each of e, and r. On other hand, WBCID have shown the best performance in terms UOI and time complexity. However, all other algorithms have scored better performance than NHR and WBCID in other criteria. Furthermore, **Appendix D** showed Tarel algorithm scored the best performance only in (e) criteria. Also this algorithm share same performance level with Kim method. However, all other algorithms have scored better performance than Tarel in other criteria. Moreover, **Appendix** E stated that GF algorithm scored the best performance in each of SSIM and UQI. However, all other algorithms have scored better performance than GF in other criteria. Tarel, Kim, WBCID, CODHWT, and Zhu have same performance value in color saturation metric (Σ). Besides, **Appendix F** exhibited that NHR algorithm scored the best performance in each of e and r. However, all other algorithms have scored better performance than NHR in other criteria. Other algorithms such as JBF and WBCID have same performance value in color saturation metric (Σ). Also, **Appendix G** presented that NHR algorithm scored the best performance in each of e r, and UQI. However, other algorithms have scored better performance than NHR in other criteria. All other algorithms such as Dehazenet, MSCNN, Zhu, Multiband, BF, WBCID, Kim, and Tarel algorithms have same performance value in color saturation criteria (Σ).

In **Appendix H**, all algorithms scored the leading performance within distinct criteria. Other algorithms such as Zhu, WBCID, and Tarel algorithms have same performance value in color saturation criteria (Σ). Finally, **Appendix I** displayed NHR algorithm scored the best performance in each of e and r. However, all other algorithms have scored better performance than NHR in other criteria. Other algorithms such as MSCNN, Zhu, Kim, and Tarel algorithms have same performance value in color saturation criteria (Σ).

V. Discussion

This study mainly aims to provide a holistic view of recent trends and issues in image dehazing. This review is also unlike other reviews because it utilizes a systematic approach (protocol) in collecting pertinent works on image dehazing. Furthermore, it offers a taxonomy of correlated literature.

Nonetheless, a noticeable leverage of developing a taxonomy for the literature exists in the research domain, particularly an emerging one. In this context, a taxonomy of the existing literature brings a well-organized approach for a series of publications. For instance, a researcher who attempts to investigate image dehazing trends may be disappointed by the huge number of designated articles for a relevant topic that do not encompass any type of structure. In this case, the researcher could fail to obtain insights into the current scenario in this field of study. Most studies approach topics from an introductory perspective, others highlight a volume of existing methods and evaluation approaches, and some offer new image dehazing algorithms and propose new metrics for the field. In addition, a taxonomy of the related literature facilitates the organization of numerous works and activities into an expressive, controllable, and well-knit scheme. Furthermore, a well-structured taxonomy is beneficial to all researchers with respectable views on the subject field in a number of ways. First, a taxonomy provides prospective guidelines of research in the field. For example, in this study, the taxonomy of

TABLE IV. Evaluation Results Based on Inhomogeneous Foggy Scene (Live)

Algorithm	e	r	Σ	HCC	SSIM	UQI	Time
Dehazenet	11.029412	1.318396	0.0016	-0.0693	0.8629	0.8337	2.4931
MSCNN	10.925098	1.4033	0.0027	0.7857	0.8982	0.9127	2.1131
Colores	11.400392	1.4669	0.0347	0.0065	0.8507	0.8309	1.3199
Zhu	11.253725	1.4080	0.0016	0.0043	0.8748	0.8997	2.4770
Multi-band	10.931765	2.8822	0.0376	-0.0331	0.5932	0.7509	0.9475
CODHWT	4.988627	1.255105	0.002373	**0.844508**	0.950891	0.945836	2.331291
Meng	14.095686	2.091165	0.034902	-0.121426	0.68192	0.698885	5.865411
Liu et al.	5.362353	1.583764	0.004549	-0.158128	0.641053	0.59172	1.171679
Berman	7.902745	**3.310823**	0.061231	-0.041326	0.55014	0.706861	11.24844
BF	8.451373	1.355124	0.0415	0.1308	0.9457	0.9842	5.4083
WBCID	16.463137	1.050864	0.0031	0.4681	**0.0529**	**0.6189**	**0.8715**
GF	7.902745	1.305597	0.0441	0.1212	0.9462	0.9892	3.8475
JBF	7.607059	1.190982	0.0345	0.1306	0.9585	0.9710	2.8108
Kim	8.991765	1.151766	0.0010	0.1106	0.9477	0.9668	1.6117
NHR	7.32	3.223612	0.1326	0.0398	0.6632	0.9246	35.8466
He et al.	6.157647	1.561816	0.0990	0.0657	0.8792	0.9880	21.0095
Tarel	**25.379608**	2.593767	**0.0001**	-0.0325	0.6911	0.8651	4.8757

image dehazing shows researchers the level of interest in developing new real-time methods; in turn, researchers could notify others about the development of image dehazing applications. Therefore, a potential direction may contribute to this area. Moreover, such an overview could facilitate the assessment of current image dehazing methods or the exchange of experiences in developing new image quality assessment methods. Meanwhile, taxonomy helps expose open issues in the available image dehazing assessment methods, that is, it outlines the articles on image dehazing into discrete classes, thereby providing a chance to investigate weaknesses and strengthens in terms of research coverage. For example, as many studies have highlighted, "to date, there is no acceptable image dehazing quality methodology." Combined with the developments of image dehazing methods in an adequate and representative sample of the literature, taxonomy also brings out several aspects of these methods, such as the execution time and accuracy of depth map estimation, which have received significant attention in the literature relative to traditional image dehazing methods. In addition, the statistics of individual categories of taxonomy highlight the environmental domains and the variety of real life applications that are based on the image dehazing concept. Nevertheless, to the best of our knowledge, most previous reviews were based on general aspects, such as categories of image dehazing algorithms. Thus, our taxonomy effectively exposes different concepts in image dehazing, such as evaluation and dataset study categories. Finally, researchers who are experts in this area can point out considerably to our taxonomy. If adopted, they can use a common language, thereby facilitating the sharing of future works and further discussions that cover areas such as development studies, new evaluation schemes, new datasets, comparative studies, and reviews on different image dehazing techniques and methods. Our study also reviews and identifies the different kinds of datasets used in the existing literature. We also illustrate different types of evaluation methods, such as objective and subjective methods, and the new evaluation metrics and methods.

However, the evaluation experiments revealed different observations. First, algorithms such as WBCID have leading performance in distinct criteria with different evaluation foggy perspectives in both examined datasets. In contrast, NHR algorithm have best performance in visibility criteria (e and r) within three foggy evaluation scenes, but only in evaluation based on RESIDE dataset. Second, the best performance for a certain algorithms cannot be achieved with more than three criteria. In other word, most of the leading algorithms have best performance in few criteria with distinct foggy evaluation scene. Third, some algorithms have shared same performance value in distinct criteria and foggy evaluation scene. Fourth, overall there is noticeable variation in the performance of each algorithm within each distinct evaluation scenario. Fifth, based on evaluation experiments in one or both datasets, there is no single algorithm have scored the best performance within all criteria as well as evaluation foggy scenes. Thus, due to performance confusion of all examined algorithms; selection of the best image dehazing algorithm is a challenging task. Therefore, our evaluation experiments confirmed the views about the selection problem that have been revealed by [6], [24], [25].

According to existing studies, the next sections describe three aspects of the literature content, namely, the motivations behind adopting image haze removal algorithms; the challenges and obstacles of developing image dehazing algorithms, evaluation methodologies, and datasets; and recommendations to mitigate such hurdles.

A. Challenges

The haze removal process and quality evaluation for degraded images are still highly challenging. Image defogging is a transdisciplinary challenge because it needs information from various aspects, such as meteorology for demonstrating mist, optical physics science for observing the manner by which light is influenced by haze, and signal processing for recouping the parameters of scenes [44]. According to investigations in the existing literature, several obstacles exist and require substantial efforts from researchers and developers to permanently align the image dehazing process with adequate restoration and enhancement results. The challenges illustrated in the literature and the citations of relevant references are discussed below. Additionally, the challenges are classified into several groups, as shown in Fig. 8.

1. Data Acquisition Challenges

Due to haze is an outdoor phenomenon, factors such as weather condition, dynamic objects, and so on have made data acquisition a hard task. This subsection presents obstacles relevant to data acquisition into image dehazing domain as follows:

a) Absence of the Haze-free Image (Ground-truth)

The assessment process for perceived image quality, especially with full reference metrics or decreased reference ones, may need two types of images, namely, hazy and haze-free images, which are taken

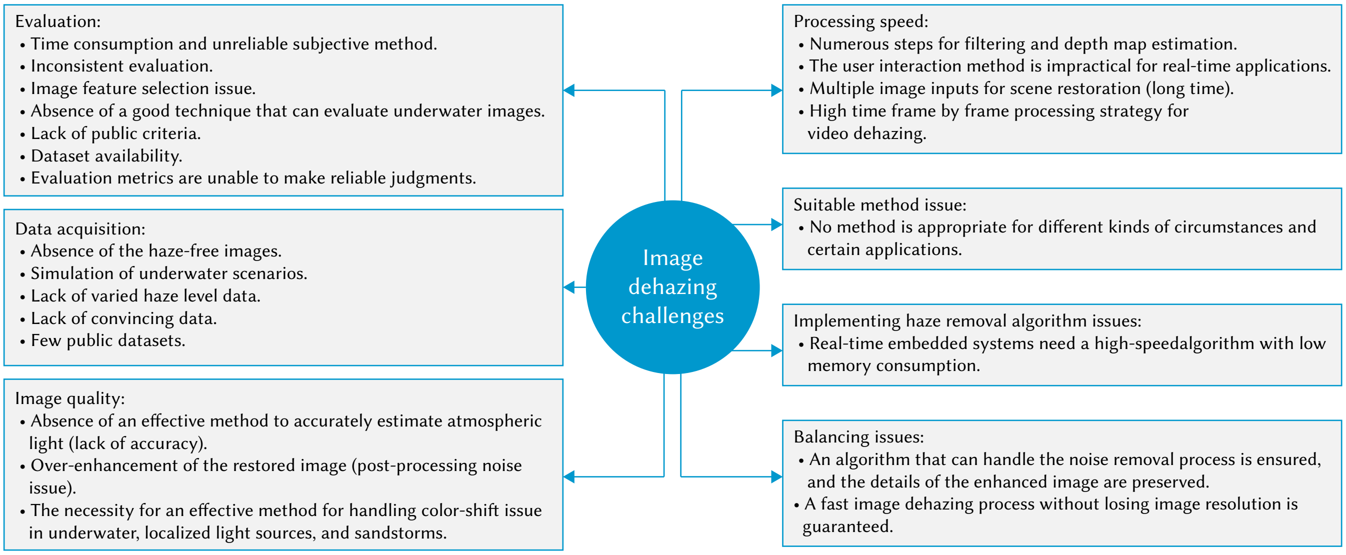

Evaluation:
• Time consumption and unreliable subjective method.
• Inconsistent evaluation.
• Image feature selection issue.
• Absence of a good technique that can evaluate underwater images.
• Lack of public criteria.
• Dataset availability.
• Evaluation metrics are unable to make reliable judgments.
Data acquisition:
• Absence of the haze-free images.
• Simulation of underwater scenarios.
• Lack of varied haze level data.
• Lack of convincing data.
• Few public datasets.
Image quality:
• Absence of an effective method to accurately estimate atmospheric light (lack of accuracy).
• Over-enhancement of the restored image (post-processing noise issue).
• The necessity for an effective method for handling color-shift issue in underwater, localized light sources, and sandstorms.
Image dehazing challenges
Processing speed:
• Numerous steps for filtering and depth map estimation.
• The user interaction method is impractical for real-time applications.
• Multiple image inputs for scene restoration (long time).
• High time frame by frame processing strategy for video dehazing.
Suitable method issue:
• No method is appropriate for different kinds of circumstances and certain applications.
Implementing haze removal algorithm issues:
• Real-time embedded systems need a high-speedalgorithm with low memory consumption.
Balancing issues:
• An algorithm that can handle the noise removal process is ensured, and the details of the enhanced image are preserved.
• A fast image dehazing process without losing image resolution is guaranteed.

Fig. 8. Categories of challenges for image dehazing.

under the same real scene settings and weather conditions [44]. Most datasets do not include images as a reference because these types of graphics are difficult to provide, thereby resulting in major difficulties in presenting an efficient evaluation process. The procedure of recording haze-free (reference) and hazy images in similar illumination conditions is still highly challenging [36],[43]-[45],[55].

b) Reproduction of Underwater Environment Characteristics

Fundamentally, light that travels through an underwater medium loses its intensity because of attenuation. Attenuation is limited to the type and number of existing particles in the turbid medium. These phenomena are due to two aspects, namely, absorption and scattering [109]. Absorption completely eliminates light beams, whereas scattering alters the course of the light spread. Simulating these phenomena is difficult because they occur due to specific particles and properties present in oceans, rivers, and lakes. Another challenge is related to the reproduction of an untouched seabed in a controlled space with specific underwater properties [46].

c) Lack of Datasets on Different Haze Levels

Image haze removal has been extensively studied, but no such image database regarding haze levels is present. Verifying the assumptions or priors that are supposed to be useful for haze removal is inconvenient for readers. Meanwhile, comparing the performance of haze removal methods, which are effective for images with different haze levels, is inappropriate. Haze level determines the amount of contrast and other details regarding a particular image [48].

d) Synthetic Image Database Issue

The FRIDA dataset [110] was based on virtual road scenes developed using computer graphics techniques. This dataset includes 66 images because of the diminished complexity level in the scenes, and some parameter settings are ineffective for real life circumstances, thereby making it less convincing for evaluation [45].

e) Lack of Benchmarked Dataset

According to [32]-[33], no public benchmark dataset on image defogging is available for comparing the performance of many enhancement algorithms.

2. Evaluation Challenges

In general, a quantitative assessment of dehazing algorithms based on a single input image is unlike other image processing methods. According to [22], several issues have been highlighted by authors in terms of the method's ability to decide on a highly enhanced image by a specific algorithm, and the answers of numerous evaluators are often inconsistent. Moreover, approving and choosing the accurate haze removal result for a particular situation is difficult. The common image quality assessment methods seldom provide solutions to these problems. Likewise, procedures that can effectively measure the quality of dehazed images using a specific algorithm are lacking [24]. Moreover, objective quality assessment methods are rarely used because they are unable to make reliable judgments [64]. The existing literature reports that no generally accepted methodology for evaluating image dehazing performance is available [36], [49], [32], [39], [22], [106]. The lack of an acceptable evaluation methodology can be classified according to following issues:

a) Subjective Evaluation Methodology Issues

Developing methodologies for evaluating such algorithms with regard to their perceptual quality is necessary. Measuring the perceptual quality of a contrast enhancement method applied to images degraded by fog is a nontrivial task, and no agreed-upon methodology currently exists [35]. The well-known solution is hand posting various degraded images and their relevant enhanced ones, which are processed by diverse algorithms, and then subjectively comparing them. However, the quantity of listed images is bounded; thus, reporting an algorithm that effectively performs in these listed images is difficult, and performances are still unknown in other cases [33]. Moreover, a subjective evaluation could include emotional responses and subjective judgments of a particular observer. Therefore, a pair of observers may end up with dissimilar or dissonant results on the same tested image [2]. Thus, human bias is allowed, and using a subjective evaluation method is expensive and time consuming, thereby making it unsuitable for real-time applications [33], [75].

b) Lack of Evaluation Consistency

Although substantial research has been conducted on image dehazing, evaluation methods presenting satisfactory results are still lacking [38]. The defogging effect assessment is difficult because the evaluation criteria for the defogging effect should be consistent with human visual perception. Image quality assessment metrics, such as mean square error and peak signal-to-noise ratio, have been widely used to assess dehazing algorithms. However, these indicators often obtain inconsistent results [39]. Moreover, comparing single indicator scores and utilizing a regression-based prediction model are inconsistent with the human visual perceptual mechanism [32], [89]. Thus, developing an evaluation method that can possibly cross over any barrier between computable assessment models and human visual perceptual mechanisms is challenging [22].

c) Appropriate Image Feature Selection

Extracting truly intrinsically salient features to define hazy images and differentiate hazy-free images from non-hazy ones is one of the evaluation challenges [32]. In addition, computational efficiency is important; thus, the features need to be immediately extracted [6].

d) Unsuitability of Evaluation indexes and Methods for Some Scenarios

According to [46], an efficient evaluation technique for underwater images is lacking. Scattering and absorption are two main issues when light travels in an underwater environment, which generates different kinds of distortions for an underwater image. One issue is color loss due to light absorption in the water. Scattering also is another issue that usually affects image details, thereby blurring image edge information and diminishing image contrast. However, utilizing (over land) atmospheric image quality assessment metrics to successfully evaluate the quality of an underwater image is difficult because of contrasting imaging concepts [40]. In [111], few metrics were developed for the evaluation of underwater images. Dehazing algorithms are aimed at achieving high image visibility (contrast) and structuring similar images. Specifically, the perceived quality of an enhanced image should match the non-hazy one (sunny day) in terms of contrast and structure. Additionally, using brightness as an evaluation metric is ineffective because the brightness of a dehazed image is different from that of a sunny one [41].

e) Lack of Public Criteria

The human visual perception itself is not a deterministic procedure. Thus, deciding the highly vital features that influence visual decision and designing corresponding evaluation metrics are difficult [39]. For the evaluation of image dehazing algorithms, the authors found that no public criterion and dataset are available for reference [33]. Moreover, no perfect criterion can effectively evaluate the quality of a perceived dehazed image [6], [8], [33].

f) Desired Dataset Availability

Different assessment techniques have been generally accepted; thus, having a highly reliable dataset is important [45]. The quality assessment for numerous algorithms has become extremely challenging due to the lack of perfect images to be used as a reference

[36], [39], [43], [44], [55], [89]. Furthermore, no public benchmark dataset for image dehazing that can be used for the evaluation process is available [32], [41].

g) Lack of Reliable Indicators

A solid image quantitative assessment metric that can viably gauge the quality of an enhanced image and the amount of information loss in restored graphics is current not available. Developing a highly reliable framework of metrics that can present a satisfactory performance for image quality assessment is a challenge because the current quantitative assessment metrics are unable to make dependable judgments [24].

h) Suitable Method Issue

Many researchers have ignored several issues in the image dehazing field. For instance, no suitable method is available for different kinds of conditions [27]. The authors reviewed some underwater image processing methods to guide other researchers in determining highly suitable techniques or methods for a particular application [31], [59]. In addition, categorized enhancement techniques based on several approaches have been used to enhance and restore hazy images and then select the appropriate algorithm for certain needs [1], for example, reviewing various methods for highlighting the suitable scheme for a driver assistance system [62].

3. Processing Speed Challenges

Methods dedicated to work in real-time applications usually need fast computation. In general, a time-consuming process is an undesirable and highly challenging problem in real-time scenarios. Many studies have been proposed to address this challenge [9], [66], [69], [86], [112]. The high computation algorithm problem could be due to the following issues.

a) Complex Computation Processing

The haze removal process becomes challenging because of unknown depths and its dependence on defined depth (transmission) maps for scenes [113]. The desired atmosphere veil should always be refined [94]. A full dehazing process consists of three complex computation steps (i.e., estimation of atmospheric light, acquisition of atmosphere veil, and restoration of a non-hazy image) [84]. The acceleration for refining transmission is a highly desirable aspect in many algorithms, such as bilateral [114], anisotropic, edge-preservation [115], and median [15] filtering, given that most image defogging algorithms need to decrease the complexity of filtration. The aforementioned algorithms are challenging to implement and apply in real-time systems [116] because they require considerable time to enhance restored images. Thus, having the minimum filtering steps is necessary for meeting real-time requirements [56], [84], [86], [94], [117], [118] restoring images without estimating airlight and transmission (depth) maps [98], [102], [103], or minimizing the time needed to calculate transmission maps [77], [69].

b) User Intervention

Depth-based methods need depth information either from known 3D models or from user interactions [119]. These types of methods are impractical to use in real-time applications because of their complexity and time-consuming nature [91]. In [120], a user must interactively register a weather-degraded image with a 3D scene model to dehaze the former. The necessary user intervention (the sky area requires to be marked out by hand) [121] and additional data for these methods make them impractical for real-time applications [96]. Many algorithms have been proposed to prevent this user interaction issue [9], [92], [93], [100], [102].

c) Multiple Image Issue

Multiple images based on the same scene were used in [3], and other ones that were taken in different weather conditions have been utilized as references for graphics that were obtained under clear weather conditions. Algorithms based on a multiple image approach are unsuitable for real-time applications [96] because of high computational complexity [122].

d) Video Processing Issue

To date, numerous efforts have been initiated to eliminate haze from single images. However, few studies have concentrated on video sequence processing. These haze removal techniques for video sequences mostly utilize a frame-by-frame strategy. In these approaches, the fundamental thought of most methodologies lies in the calculation of depth maps for degraded scenes through the use of multiple images under various climate conditions [80]. Moreover, the high time complexity of video dehazing occurs when utilizing a frame-by-frame strategy. Many methods have been designed with different strategies to prevent the time-consuming processing of the frame-by-frame strategy for achieving real-time video dehazing to address this efficiency issue [81], [82], [123], [124].

However, a fast execution time is an essential step for certain real-time video or image applications implemented in embedded systems. For example, considering a 30 fps video, the processing time of one frame in such a video must be no more than 0.03 s to meet real-time application requirements [72], [73], [85], [86], [97]. Therefore, an enhancement or restoration algorithm that can process 30 fps is suitable for real-time image applications.

4. Image Quality Challenges

a) Estimation Accuracy Issue

Airlight must be refined in terms of time, except for some regions wherein the depth map randomly changes because atmosphere veil essentially relies on the information of scene depth [94]. One of the crucial steps in image defogging is to provide highly accurate restored images on the basis of the accurate estimation of transmission maps. Despite the development thus far, an efficient method that accurately estimates the global atmospheric, which is a highly crucial part in the quality of image restoration, is lacking [68], [125], [90], [95].

b) Over-enhancement Issue

Many image dehazing methods, such as the DCP, based on a single image input have been investigated in the literature. The authors in [10] described the notion of dark channels on the basis of the thought that "in a clear day image, except for the sky regions, the intensity of each pixel will be close to zero at least in one color channel." This statistical observation is called DCP [5]. However, the image resulting from the restoration process exhibits missing details and suffers from unnatural coloring [58], and some haze remains at the edges of the images [83]. Specifically, the two main issues in the methods according to the DCP scheme are color distortion and generated halo artifacts in restored images [70]. Moreover, DCP cannot efficiently work with scenes that contain sky regions and white objects, under which it leads to a severe color distortion or blocking effect in restored images [58]. Thus, post-refinement processing is required to preserve image edges and efficiently restore color. Many methods have been proposed to effectively handle the defect generation of halo artifacts and color distortion to address the image darkening issue resulting from the restoration process [10], [50], [69], [86], [112], [126].

c) Color Distortion Issue

Color change is one of the main distortion issues for underwater images. The amount of color distortion increases according to the variation of the attenuation degree that the traveling light is exposed to [92]. Apart from the underwater scene, another scenario for color distortion is the localized light sources, which usually occur when car drivers turn on the headlights of their cars and streetlights are activated, thereby causing localized light in images that have been taken in these circumstances. Sandstorm is another weather circumstance that is normally experienced through driving in some areas. During a sandstorm, the atmospheric sand can absorb particular parts of a spectrum, thus producing color-shift problems in the taken image. In summary, the common up-to-date restoration algorithms are incapable of efficiently dealing with hazy images that feature color-shift problems or localized light sources [15].

5. Balancing the Noise Removal Process and Preserving Details of the Enhanced Image

Apart from the enhancement issues, one of the image dehazing obstacles is to guarantee the obtainment of a high-quality restored image without any image information loss. Balancing the decreasing defects resulting from the dehazing process while keeping the proper quality of restored images is difficult [23]. For example, the contrast of road scene images is considerably enhanced through the use of common image dehazing methods [75]. Moreover, a histogram equalization is used to increase the contrast of foggy images, but the quality is still extremely poor because noise increases as the image contrast improves [127]. In summary, the issue of balancing could occur between certain criteria, such as visibility and color fidelity [97], wherein an over-saturated image may present a high contrast gain but show a large number of saturated pixels [9].

6. Balancing the Quality and Speed of Image Restoration

Achieving a fast single image dehazing has been a challenging issue in many fields, such as real-time applications [89]. The existing literature shows that balancing computation time and quality of restored images is still an open issue [96] because providing a process that can offer a short time and present no image resolution loss (image details) for the image enhancement process is difficult [128]. Thus, restoring images to their natural conditions and ensuring the balance between the speed of image restoration and perceived quality are vital steps in the image dehazing process [69], [86], [116].

7. Issues in Implementing Haze Removal Algorithm

Most multimedia applications that require real-time processing currently rely on multi-core embedded systems because of their extraordinary data rate processing capabilities [71]. Real-time applications that support embedded systems usually need an algorithm that can handle several requirements, such as low memory consumption and fast processing (speed), of real-time scenarios. This scenario is challenging through the implementation of haze removal algorithms for image sequences on embedded systems [5], [71], [81], [129].

B. Motivations

Using image dehazing technology in various application domains and scenarios has numerous benefits. Thus, researchers are motivated to further improve image dehazing technology. This section demonstrates the multiple advantages revealed in the existing literature, which are classified in particular groups with corresponding reference citations (see Fig. 9).

1. Image Retrieval Benefits

In such situations, the intensity of reflected light from any scene point is usually attenuated as it travels to the camera device. In addition, airlight acts as the main source of illumination for all objects in the scope of the scene [130]. The major drawback of the abovementioned situations is the reduced image visibility, thus resulting in considerable hindrances in various computer vision applications, such as image retrieval, photography restoration, and scene analysis [122], [128]. In general, image restoration methods have been made to overcome the defects of the quality of hazy images [46]. The notable feature of image dehazing is the restoration of image details, such as color [60] and contrast [86].

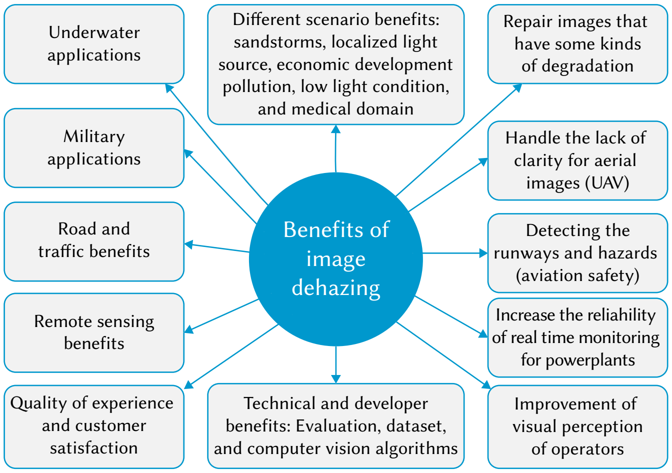

Benefits of image dehazing
Underwater applications
Different scenario benefits: sandstorms, localized light source, economic development pollution, low light condition, and medical domain
Repair images that have some kinds of degradation
Military applications
Handle the lack of clarity for aerial images (UAV)
Road and traffic benefits
Detecting the runways and hazards (aviation safety)
Remote sensing benefits
Increase the reliability of real time monitoring for powerplants
Quality of experience and customer satisfaction
Technical and developer benefits: Evaluation, dataset, and computer vision algorithms
Improvement of visual perception of operators

Fig. 9. Image dehazing motivation categories.

2. Military Applications

The quality enhancement of images obtained in foggy circumstances is a highly notable research area in military and civil applications [131], [42], [106]. UAVs are utilized for detecting (reconnaissance) and attacking intruded ground targets. The main drawback for this category of UAVs is the inefficiency of the object detection process, especially under turbid weather conditions, in which objects could hit the UAVs [132]. Thus, increasing the accuracy of detection is necessary to improve reliability.

3. Underwater Application Benefits

Underwater images suffer from tough color distortion and noise from artificial light sources, thereby resulting in image blurring and haziness [60]. In recent years, different underwater applications have been widely developed, with examples including documentation support and aircraft accidents. Fundamentally, remote-operated vehicles controlled by human experts are used to perform specific interventions. Highly advanced vehicles, namely, intervention autonomous underwater vehicles, have been recently developed, as described in [133]. One of the main disadvantages of AUVs is the necessity to understand the hostile environment to detect and recognize objects within obscured scenes. This condition is the motivation behind many studies that examined underwater image degradation to restore and enhance the visibility of underwater scenes [31]. Thus, the process of restoring the color of underwater degraded images will provide clear images for many underwater applications, such as the following.

- Monitoring marine biodiversity (exploration) [59].
- Underwater rescue (safety) [59].
- Detecting underwater pipeline leaks (pollution reduction) [59].
- Underwater microscopic detection [60].
- Terrain scanning (accurate terrain classification) [58], [134].
- Mine detection (safety of vessels and human lives) [60].
- Telecommunication cables (maintenance and tracking) [60].
- Coral image classification [135].
- Documenting underwater archaeology (shipwrecks) [136].

4. Road and Traffic Benefits

Fog is one of the critical factors of road accidents in bad weather conditions [27], [79]. In addition, road safety is considered one of the major issues in the transportation strategy of the European Union Commission [137]. In general, fog lessens the visibility of a scene, thereby affecting the visual quality of images [13]. For example, drivers cannot distinguish a road scene in foggy weather [26]. The demand for cameras and intelligent surveillance systems that execute real-time recording and monitor private or public areas is currently increasing. For example, consider a vehicle dashboard camera. Such camera records road situations in real time during driving, and the recorded videos can provide information for automatic license plate recognition [138] or data on crash-related events. Moreover, such information is highly beneficial for police investigating and resolving road incidents and traffic violations [137]. In intelligent transportation systems, cameras keep track of road or street scenes to detect traffic flow or identify cars for specific applications, such as vehicle identity checks. Therefore, video quality plays a critical role in such applications. However, unfavorable weather conditions or poor atmospheric light result in inadequate visibility in imaging systems and lead to the production of blurred video images, thereby rendering the recorded videos ineffectual. Defogging is a vital preprocessing technique for object detection in computer vision-based systems, and it has been widely used in outdoor surveillance system applications [85]. Furthermore, driver assistance systems take advantage of the increasing accuracy of detection and feature extraction through the availability of an image fog removal tool for the following processes: improvement of road-marking feature extraction and camera-based obstacle and circular road-sign detection [75], [76].

5. Different Enhancement Scenario Benefits

Different circumstances require image dehazing techniques to enhance degraded images. For instance, the manner by which image dehazing principles are used to decrease the color shift problem or distortion for two scenarios (i.e., localized light sources in images captured when drivers turn on their headlights and images taken through sandstorm weather) was observed [139]. Moreover, the authors considered image dehazing in videos taken under low lighting conditions to handle several particular issues, such as poor visibility and contrast [140]. The authors in [66] mentioned that fog is worsening in China as the economy develops. Most image recognition systems are suitable for normal weather, but the applications based on the restoration of degraded images are highly valuable.

6. Medical Domain

Recently, a new direction for image dehazing domain have been appeared where image dehazing theory can be applied for different case studies of medical area. For instance, the premature infants' retinal images are generally of lower visibility compared to adult retinal images, affecting the quality of diagnosis. Authors in [141] studied some image dehazing methods from general outdoor scenes and proposed an image restoration scheme for neonatal retinal images. Also, Medical X-ray image its quality digressed because of the interferences caused by the human body structure, equipments, and environmental factors. Authors in [142] ,verified that the X-ray image degradation caused by the X-ray scattering which is similar as the haze scattering.by applying dark channel prior method this challenge can be solved.

7. Consumer Market Benefits

Consumers prefer non-hazy images with high visibility details when shooting target objects. Consequently, image editing software or cameras that can restore scene details in hazy or foggy weather are highly beneficial for consumer marketplaces, and camera and camcorders increase customer satisfaction and reliability [55], [68], [91], [143]. Furthermore, televised transmissions of outdoor sports events, such as cross-country skiing or ski jumping, during hazy weather can seriously affect the quality of experience of television audiences [35].

8. Technical and Developer Benefits

a) Evaluation Benefits

According to the comparison result of various graphics, the quality of several enhanced images is poorer than that of hazy ones [36]. The manner of effectively comparing the performance of image dehazing becomes a novel task with the advancement of haze removal techniques in the past few years [32]. Thus, the evaluation process can present such an advantage in terms of measuring the effectiveness of enhancement quality for a particular algorithm against other ones.

b) Dataset Benefits

With regard to the evaluation of several algorithms and the development of a new one, providing clear images and supplementary datasets is highly essential to perform previous processes. Furthermore, a successful evaluation is obtained when we gauge the enhancement in processed (hazy) and reference (non-hazy) images [43], [44]. However, the formation of an outdoor scene is highly complex and relies on different atmospheric circumstances, such as mist, clear air, and fog. Thus, a large number of images are essential to study the complete variations of scene appearances. Datasets are used to reveal the importance of supporting the process of developing new algorithms for enhancing the visibility of degraded images for computer vision applications [45], [47], [48].

c) Computer Vision Algorithm Benefits

Many computer vision algorithms [144], such as image segmentation [145], annotation [146], and matting [147], are used when recovering a haze-free image from a bad one. On this basis, research on image dehazing has important and realistic importance [24], [125].

9. Remote Sensing (Measurement) Benefits

The evaluation of haze may facilitate recognition for images with extremely poor visibility to increase the reliability of remote-sensing image analysis. Remote sensing images offer substantial information about geographic and spatial areas and have been extensively used in hydrology, forestry [148], and weather forecasting [104]. However, all images that require remote sensing analysis easily suffer from the effects of haze on the visibility of specific scenes, thereby decreasing the value of remote sensing applications to a boundless level [42], [88], [98], [131].

10. Improvement of Visual Perception of Human Operators

In hazy weather, the severe degradation of the information captured by optical sensors usually occurs because of the scattering of atmospheric particles. Specifically, the attenuation of atmospheric light decreases the contrast and fidelity of images, thereby directly affecting the visual perception of the human operator vision system [77], [149]. Thus, studying the methods of image dehazing is necessary [86].

11. Power Plant Monitoring Benefits

Zones (i.e., near mountains) around plants frequently experience hazy weather due to their locations, thereby affecting the visibility of video monitoring. In addition, this type of noise affects the work of personnel, which involves extracting important information from certain videos, particularly in terms of line monitoring stations; brings enormous hazards to the information analysis process, and calls for early warning when wire line information is difficult to distinguish. Power lines should eliminate fog effects to increase the visibility of surveillance videos.

Furthermore, solving the foggy video problem is necessary to increase the reliability of power plant monitoring in real time [77].

12. Detecting Runways and Hazards (Aviation Safety)

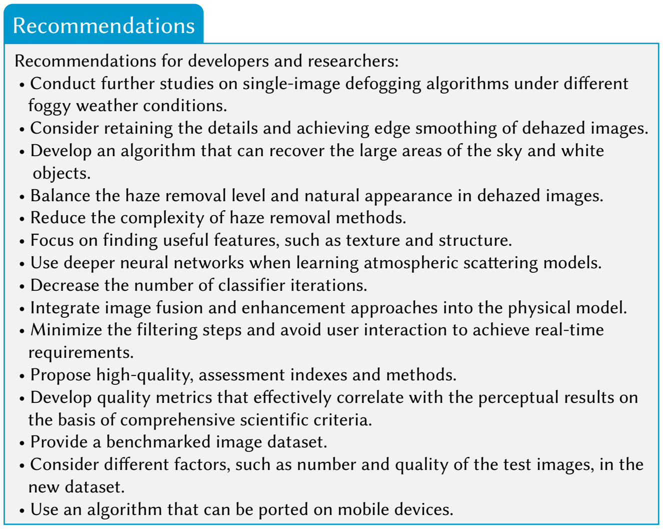
The visibility of a scene decreases as the density of fog increases, thereby causing difficulties in aircraft take-off and landing (runway detection) [126], [150]. Thus, improving visibility and making images that are pleasing are beneficial for various applications, such as runway hazard detection [5].

13. UAV Benefits

Suspending particles usually produced by hazy weather easily affect the image formation process in UAV images [82], [63]. Hazy or foggy circumstances extremely diminish the visibility for the UAV imaging system, thereby resulting in a decreased reliability for UAVs. With regard to the extraction of important image features and target detection, low contrast and visibility are undesirable [151]. Thus, the aerial image defogging process in fog conditions is a highly beneficial aspect [152].

C. Recommendations

We also summarize many notable recommendations in the existing literature to support the image dehazing community in terms of diminished challenges and facilitate the development of image dehazing techniques (see Fig. 10).



Recommendations

Recommendations for developers and researchers:
- Conduct further studies on single-image defogging algorithms under different foggy weather conditions.
- Consider retaining the details and achieving edge smoothing of dehazed images.
- Develop an algorithm that can recover the large areas of the sky and white objects.
- Balance the haze removal level and natural appearance in dehazed images.
- Reduce the complexity of haze removal methods.
- Focus on finding useful features, such as texture and structure.
- Use deeper neural networks when learning atmospheric scattering models.
- Decrease the number of classifier iterations.
- Integrate image fusion and enhancement approaches into the physical model.
- Minimize the filtering steps and avoid user interaction to achieve real-time requirements.
- Propose high-quality, assessment indexes and methods.
- Develop quality metrics that effectively correlate with the perceptual results on the basis of comprehensive scientific criteria.
- Provide a benchmarked image dataset.
- Consider different factors, such as number and quality of the test images, in the new dataset.
- Use an algorithm that can be ported on mobile devices.

Fig. 10. Image dehazing recommendation categories.

1. Recommendations of Developers and Researchers

a) Image Dehazing Methods

Conducting further studies on single-image defogging algorithms that can adaptively enhance foggy images acquired under different foggy weather conditions [6], [46], [64], such as night, dense, and inhomogeneous foggy weather; night conditions [74], sandstorms [50], effects of shadow, and unwanted light [126], is recommended. Several proposed algorithms suffer from various post-processing effects, such as dimness at the edges of dehazed images; thus, preserving the details, achieving edge smoothing for dehazed images [86], and performing additional improvements on existing dehazing methods in the future are recommended [116]. Notably, several methods fail to recover the large areas of the sky and white objects [153], [154] of degraded images. Thus, considering these issues when developing a particular method is important. Moreover, future dehazing algorithms must possess a balance between the natural appearance and the effects of the dehazing process for a specific image [36], [64].

In some existing defogging algorithms, parameters need to be manually set. Although an excellent performance can be obtained by constantly adjusting parameters, the result is unrealistic in real-time applications [6]. At present, most video dehazing processes are improvements of single image dehazing methods and usually contain complex data processing algorithms. These complex operations often require a long processing time [8]. However, most existing image dehazing algorithms have high time and space complexity. Most desirable scenarios in real-time applications provide an automatic and adaptive processing for needed images. Thus, reducing the complexity of haze removal methods [64], constantly adjusting performance parameters [6] or minimum user interaction [62], and concurrently processing a large number of videos such that atmospheric light estimation can be shared and coordinated between different videos [123], are recommended.

Emphasizing the search for beneficial features, such as texture and structure [125], and recovering degraded images from as few features as possible are recommended to establish a highly powerful neural network model for single-image dehazing [22]. Moreover, using deeper neural networks in the learning of atmospheric scattering models is suggested; in this case, an end-to-end mapping between hazy and haze-free images can be directly optimized without any need for estimation of medium transmission [68]. The number of classifier iterations should be minimized as much as possible to obtain the final result [32].

Apart from the widely used atmospheric scattering model, degradation models, such as the dual-color atmospheric scattering and atmospheric transfer function, are currently available. However, none of these models can accurately describe the phenomenon of haze degradation. Therefore, exploring some cues that have been obtained from research results of modern atmospheric optics (study of comprehensive degradation models), is necessary [8].

Many image enhancement methods have been developed on the basis of the human vision system. These methods can rapidly and accurately estimate image brightness and maintain true color. Image fusion methods can determine or obtain effective information from different source images. Thus, integrating image fusion and enhancement approaches into physical models is recommended [8].

All existing video defogging algorithms focus on surveillance scenes. No effective video defogging algorithm for a scene with a moving camera is available. The color shift problem also needs to be overcome in further studies [6]. Evidently, de-weathering based on multiple images and user interaction is unsuitable for driver assistance systems. A fully automated system is highly recommended for real-time image dehazing scenarios [62]. Furthermore, some algorithms have been recommended in the implementation of particular real-time applications [5], [66], [69], [87], [90], [93], [100], [112], [131], [124].

b) Image Quality Assessment

Given that previous objective assessment results are evidently inconsistent with subjective ones, directly applying them to evaluate different defogging algorithms is difficult because they are unable to make reliable judgments [6], [64]. Thus, an effective quality assessment index or method also needs to be proposed [6], [24]. At present, research on the quality assessment of dehazed images still requires further development, and the evaluation indexes are mainly concentrated on image clarity, contrast, color, and structural information and lack comprehensive scientific criteria. Thus, designing a special image quality assessment mechanism is necessary [8]. Furthermore, substantial work should be conducted to develop quality metrics that effectively correlate with perceptual results [35],[44]. The study recommends that the statistics of natural scenes in addition to the distortion of particular features be combined to generate a highly delicate objective image quality assessment method in the future [36].

c) Dataset Recommendations

Developers and researchers, such as [43], [44], [46], [55] recommend using several datasets to develop new fog removal algorithms and image quality assessment methods. However, the additional noise is a critical issue in the dehazing process and image quality assessment. Moreover, the noise in common datasets utilized for image quality assessment is artificially added. These datasets are unable to precisely reflect the real complex noise in normal hazy images. Therefore, creating a public benchmark dataset is recommended for image dehazing [22]. Furthermore, different factors, such as increasing the number and quality of test images, should be considered in new datasets [35], and hazy image datasets should include large bright regions that usually exist in natural scenes [32].

d) Cost Efficient Solution

According to [72], [73], enhancement algorithms can work in real-time environments, such as the pre-processing stages for several real life applications (e.g., traffic surveillance systems, basic image dehazing, and driving assistance systems). Additionally, these algorithms can be implemented on mobile devices. Thus, developers recommend using the aforementioned algorithms to support users with minimal cost and efficient solution for image visibility restoration in various driving scenarios.

VI. Limitations

The number and identity of the source databases are eminent limitations in our study because the process of searching related articles was based on three search engines of online databases. Nevertheless, the designated databases are reliable and provide relevant articles. In addition, the rapid development in the area affected the timeline of this survey. According to the time limitation, relevant studies do not necessarily cover the entire picture about trends, development, and effects of this area. Consequently, our study barely illustrates the number of responses from the image dehazing community to the area, which is the main target of this study.

VII. Conclusion

In the past decades, researchers have drawn notable attention to the image dehazing development, thereby making the image dehazing technology one of the major research topics. In this context, no clear boundaries have been observed in the development of this field. Thus, further study is necessary to provide a holistic view and track this research line. Our study attempts to provide an extensive view and deep understanding by reviewing and classifying the highly pertinent literature. Consequently, this study maps the final set of relevant articles in three main categories, namely, studies conducted on image dehazing, reviews and surveys, and real-time scenario-based development. Apart from providing an intensive investigation into the existing literature, the three main classes are divided into subcategories, such as comparative study, various types of evaluation methods, datasets, review articles conducted in general or supported specific scenarios, and evaluation criteria types that have been used to measure the efficiency of certain algorithms. In addition, further details, such as the challenges and obstacles in the image dehazing community, the relevant motivations behind holding a particular image dehazing study, notable recommendations stated by other researchers to mitigate existing hindrances, and various datasets that have been used to support evaluation methodologies and algorithm development processes, are presented through intensive search and analysis of the final set of articles in distinct forms. On the one hand, researchers have paid great attention to the development of real-time image dehazing algorithms. On the other hand, the existing literature reveals little concern about improving the evaluation procedure for certain image dehazing algorithms and handling related issues. Thus, the image dehazing community should exert substantial efforts toward the development of new evaluation methodologies and resolving the obstacles in the evaluation process. Finally, our systematic review will help researchers track the critical issues regarding image dehazing, thereby extending and drawing further research directions.

Appendix

Appendix A. Dataset Statistics

Ref	Dataset	Over-land	Over-water	Underwater	real	Synthesis	Indoor	Outdoor	Source
[55]	30 images based on five scenes	×	×	✓	×	×	×	×	[55]
[44]	CHIC (Color Hazy Image For Comparison) dataset = 9 images (publicly available)	✓	×	×	×	✓	✓	×	http://chic.u-bourgogne.fr
[48]	More than 3464 images	✓	×	×	✓	×	×	✓	[48]
[47]	SAMEER-TU Database = 5390 images	✓	×	×	✓	×	×	✓	[47]
[45]	D-HAZY dataset = 1400+ pairs of images	✓	×	×	×	✓	✓	×	[45]
[43]	Large-volume road scene dataset = 2000 images	✓	×	×	×	✓	×	×	[43]
[40]	Underwater dataset = 87 images	×	×	✓	×	✓	×	×	[40]
[87]	100 images	✓	×	×	✓	×	×	✓	[87]
[68]	12 pairs of stereo images collected from the Middlebury Stereo Datasets (publicly available)	✓	×	×	×	✓	✓	×	http://vision.middlebury.edu/stereo/data/
[97]	Two weather degraded videos, namely, "Riverside" and "Road View" (publicly available)	✓	×	×	✓	×	×	✓	http://mcl.korea.ac.kr/projects/dehazing/

Ref	Dataset	Over-land	Over-water	Underwater	real	Synthesis	Indoor	Outdoor	Source
[96]	IV-M dataset = 24 images	✓	×	×	✓	×	×	✓	[96]
[155]	Multiple real-world foggy image dataset (MRFID)= 200 clear images and each with four Corresponding foggy images of different densities. DMRFIs= 12,800 defogged images	✓	×	×	✓	×	×	✓	http://www.vistalab.ac.cn/MRFID-for-defoggin
[156]	Synthetic haze removing quality (SHRQ) database=675	✓	×	×	×	✓	✓	✓	[156]
[157]	22 ✓airs of hazy images and haze-free images (ground truth)	✓	×	×	×	✓	✓	×	[157]
[25]	REalistic Single-Image DEhazing (RESIDE)= 13, 990 synthetic hazy	✓	×	×	×	✓	✓	✓	https://sites.google.com/view/reside-dehaze-datasets/reside-standard?authuser=0
[158]	Overall dehazing quality (DHQ)=1,750 images	✓	×	×	✓	×	×	✓	[158]
[159]	Non-homogeneous realistic dataset NH-HAZE= 55 scenes	✓	×	×	×	✓	×	✓	https://data.vision.ee.ethz.ch/cvl/ntire20/nh-haze/
[160]	Vehicles Small Object Dataset (VSOD)	✓	×	×	×	✓	×	✓	[160]
[161]	Dense-Haze dataset =33 pairs of images	✓	×	×	×	✓	×	✓	[161]
[162]	More than 1000	✓	×	×	✓	×	✓	✓	[162]
[163]	Visibility Range Haze Simulation(VRHAZE) =8 pairs images	✓	×	×	×	✓	×	✓	[163]
[164]	57 image pairs	×	×	✓	✓	×	×	×	http://csms.haifa.ac.il/profiles/tTreibitz/datasets/ambient_forwardlooking/index.html
[165]	U45= 45 images	×	×	✓	✓	×	×	×	https://github.com/IPNUISTlegal/underwater-test-dataset-U45-
[166]	Underwater Image Enhancement Benchmark (UIEB) =950 images	×	×	✓	✓	×	×	×	https://li-chongyi.github.io/proj_benchmark.html
[167]	Over-wate Haze=4531 images	×	✓	×	✓	×	×	✓	[167]
[168]	I-HAZE= 35 image pairs	✓	×	×	×	✓	✓	×	[168]
[169]	O-HAZE= 45 image pairs	✓	×	×	×	✓	×	✓	https://data.vision.ee.ethz.ch/cvl/ntire18//o-haze/
[170]	20550 images	✓	×	×	×	✓	×	✓	[170]
[171]	CHIC (Color Hazy Images for Comparison) = two indoor and two outdoor scenes	✓	×	×	×	✓	✓	×	http://chic.u-bourgogne.fr/
[172]	LIVE Image Defogging Database=1100 images	✓	×	×	✓	×	×	✓	https://live.ece.utexas.edu/research/fog/fade_defade.html

Appendix B. Critical Analysis of Real-time Image Dehazing Algorithms

	Approach					Evaluation		Data type				
Ref	Image restoration	Image enhancement	Image fusion	Hybrid	Technique	Subjective	Objective	Image	video	Image and video	Scene type	Application support
[173]	×	×	×	✓	Multi-band decomposition	✓	✓	✓	×	×	General	Robot Vision
[174]	✓	×	×	×	CONVEX OPTIMIZATION	×	✓	✓	×	×	General	Not specified
[175]	✓	×	×	×	Boundary Constraint and Contextual Regularization	×	✓	✓	×	×	General	Not specified
[176]	×	×	×	✓	open dark channel and Wavelet	×	✓	✓	×	×	General	Not specified
[177]	✓	×	×	×	non-local prior	✓	✓	✓	×	×	General	Not specified
[178]	×	×	×	✓	Bilateral Filter	✓	×	✓	×	×	General	Not specified
[179]	×	×	×	✓	White Balance and image decomposition	✓	✓	✓	×	×	General	Not specified
[180]	×	✓	×	×	Guided Image Filtering	×	✓	✓	×	×	General	Not specified
[181]	×	✓	×	×	Guided joint bilateral filter	×	✓	×	×	✓	General	Not specified
[182]	✓	×	×	×	linear combination of the direct transmission, airlight and glow	×	✓	✓	×	×	Night-time	Not specified
[183]	×	✓	×	×	median filter	✓	✓	✓	×	×	General	lane-marking and obstacle detection
[112]	✓	×	×	×	DCP and GIR filter	×	✓	✓	×	×	General	Not specified
[50]	✓	×	×	×	HSV color space	×	✓	✓	×	×	General	Not specified
[71]	✓	×	×	×	DCP	×	✓	×	×	✓	General	Driver assistance system
[77]	✓	×	×	×	DCP	×	✓	×	✓	×	General	Power station monitoring
[78]	✓	×	×	×	Locally adaptive Wiener defogging	×	✓	×	✓	×	General	Optical system for observing targets
[87]	×	×	×	✓	Fusion weighting scheme and atmospheric light	✓	✓	×	×	✓	General	Not specified
[93]	✓	×	×	×	DCP	×	✓	✓	×	×	General	Not specified
[100]	×	×	×	✓	Retinex based and DCP	×	✓	✓	×	×	General	Not specified
[79]	×	✓	×	×	CLAHE	×	✓	×	✓	×	General	Real-time video surveillance system
[74]	×	✓	×	×	Histogram equalization	×	✓	✓	×	×	General	Road edge detection and road obstacle detection
[98]	×	×	✓	×	Per-pixel strategy	×	✓	✓	×	×	General	Not specified
[88]	✓	×	×	×	Boundary constraints and bilateral filtering	×	✓	×	×	✓	General	Not specified
[72]	✓	×	×	×	New mathematical model	×	✓	✓	×	×	General	Driver assistance system
[73]	✓	×	×	×	New mathematical model	×	✓	✓	×	×	Daytime	Driver assistance system
[90]	✓	×	×	×	DCP and median DCP (MDCP)	×	✓	×	✓	×	General	Not specified
[75]	✓	×	×	×	DCP	×	✓	✓	×	×	General	Road marking feature extraction and road sign detection

	Approach					Evaluation		Data type				
Ref	Image restoration	Image enhancement	Image fusion	Hybrid	Technique	Subjective	Objective	Image	video	Image and video	Scene type	Application support
[94]	✓	×	×	×	Bilateral and DCP guided filters	×	✓	✓	×	×	General	Not specified
[116]	✓	×	×	×	Linear transformation	×	✓	✓	×	×	General	Not specified
[154]	×	×	×	✓	Joint LLSURE	×	✓	✓	×	×	General	Not specified
[97]	×	✓	×	×	Gamma correction	×	✓	×	×	✓	General	Not specified
[56]	✓	×	×	×	DCP (guided filter)	×	✓	✓	×	×	General	Not specified
[95]	✓	×	×	×	DCP and bilateral filters	✓	✓	✓	×	×	General	Not specified
[81]	✓	×	×	×	DCP	×	✓	✓	×	×	General	embedded systems
[82]	✓	×	×	×	DCP	×	✓	×	×	✓	General	Unmanned aerial vehicle (UAV)
[134]	✓	×	×	×	HRNFP	×	✓	✓	×	×	General	Not specified
[150]	✓	×	×	×	DCP	×	✓	✓	×	×	Sky	Not specified
[91]	×	×	×	✓	DCP and multi-scale retinex	×	✓	✓	×	×	General	Not specified
[83]	✓	×	×	×	MDCP	×	✓	✓	×	×	General	Car vision systems
[84]	✓	×	×	×	CABFD	×	✓	✓	×	×	General	Not specified
[76]	✓	×	×	×	Flat-world assumption	×	✓	✓	×	×	Daytime	Road marking, road sign, and road obstacle detection
[184]	✓	×	×	×	DCP and fast Fourier transform	×	✓	✓	×	×	General	Not specified
[102]	×	×	×	✓	DCP and infrared-blue light intensity difference factor	×	✓	✓	×	×	General	Mobile cloud of smart city
[96]	✓	×	×	×	DCP	✓	✓	✓	×	×	Sky	Not specified
[128]	✓	×	×	×	Adaptive DCP	×	✓	✓	×	×	Sky	Not specified
[101]	×	×	×	✓	Digital total variation (TV) filter with color transfer (DTVFCT)	×	✓	✓	×	×	General	Not specified
[105]	✓	×	×	×	Machine learning	×	✓	✓	×	×	General	Not specified
[99]	×	×	✓	×	White balance and a contrast enhancing procedure	×	✓	✓	×	×	General	Not specified
[57]	✓	×	×	×	Color ellipsoid prior	×	✓	✓	×	×	General	Not specified
[103]	×	×	×	✓	DCP and reliability guided fusion	×	✓	✓	×	×	General	Not specified
[89]	✓	×	×	×	Mean filter (DCP)	×	✓	×	×	✓	Sky	Not specified
[92]	✓	×	×	×	Joint trigonometric filter	×	✓	✓	×	×	General	Not specified
[80]	×	×	×	✓	DCP and multiscale retinex	×	✓	×	✓	×	General	Surveillance camera system
[106]	✓	×	×	×	Machine learning	×	✓	✓	×	×	General	Not specified
[104]	✓	×	×	×	DCP and histogram-based S-shaped transfer mapping	×	✓	✓	×	×	General	Not specified
[85]	✓	×	×	×	DCP	×	✓	×	×	✓	General	Not specified
[185]	×	✓	×	×	gamma-correction operations	✓	✓	✓	×	×	General	Not specified

	Approach					Evaluation		Data type				
Ref	Image restoration	Image enhancement	Image fusion	Hybrid	Technique	Subjective	Objective	Image	video	Image and video	Scene type	Application support
[186]	✓	×	×	×	DCP	×	✓	✓	×	×	General	Agriculture
[187]	×	×	×	✓	local Laplacian filtering and DCP	✓	✓	✓	×	×	Sky	Not specified
[188]	✓	×	×	×	Fusion of Luminance and Dark Channel Prior (F-LDCP)	×	✓	✓	×	×	Sky	Not specified
[189]	✓	×	×	×	simple radiographic scattering model	×	✓	✓	×	×	x-ray industrial objects	nondestructive testing (NDT)
[190]	×	✓	×	×	multi-scale retinex	×	✓	✓	×	×	General	Not specified
[191]	✓	×	×	×	Retinex and DCP	✓	✓	✓	×	×	Dark	Not specified
[192]	✓	×	×	×	Machine learning	✓	✓	✓	×	×	General	Not specified
[193]	✓	×	×	×	image decomposition	✓	✓	✓	×	×	Dark	Not specified
[194]	✓	×	×	×	Machine learning	✓	✓	✓	×	×	General	Not specified
[195]	✓	×	×	×	Machine learning	×	✓	✓	×	×	Sky	railway industry
[196]	✓	×	×	×	DCP	×	×	✓	×	×	General	Steganography
[197]	✓	×	×	×	Machine learning	✓	✓	✓	×	×	General	UAV-based railway
[198]	✓	×	×	×	Machine learning	×	✓	✓	×	×	Inhomogeneous	Not specified
[199]	✓	×	×	×	DCP	✓	✓	✓	×	×	General	Not specified
[200]	✓	×	×	×	Machine learning	×	✓	✓	×	×	General	Not specified
[201]	✓	×	×	×	DCP	×	✓	×	×	✓	General	Measurement of vehicle safe distance
[202]	✓	×	×	×	n/a	×	×	✓	×	×	n/a	Biometric
[203]	✓	×	×	×	Machine learning	✓	✓	×	×	✓	General	Safe Autonomous Driving
[204]	×	✓	×	×	Retinex	✓	✓	✓	×	×	General	Not specified
[205]	✓	×	×	×	haze-line	×	×	✓	×	×	n/a	TV industry
[206]	✓	×	×	×	DCP	×	✓	✓	×	×	General	Agriculture
[207]	✓	×	×	×	Machine learning	×	✓	✓	×	×	Sky	Not specified

Appendix C. Evaluation Results Based on Homogeneous Foggy Scene (Live)

Algorithm	e	r	Σ	HCC	SSIM	UQI	Time
Dehazenet	4.1588	1.4960	0.0001	0.0413	0.8602	0.8410	2.5464
MSCNN	4.1538	1.5652	0.0003	0.088498	0.860096	0.8794	1.8726
Colores	8.0088	1.6923	0.0009	-0.2389	0.758971	0.723296	2.243798
Zhu	5.5400	1.5027	0.00001	-0.0139	0.847028	0.829924	2.298305
Multi-band	15.0721	2.9541	0.0005	-0.2509	**0.576338**	0.703543	0.836256
CODHWT	2.989583	1.361311	0.000051	**0.424145**	0.936578	0.934878	1.809423
Meng	10.23	2.295751	0.005086	-0.327609	0.636078	0.674045	4.793871
Liu	6.318333	1.744583	0.00019	-0.379144	0.638409	0.584564	0.962875
Berman	6.968333	2.349564	0.001553	-0.145496	0.68528	0.760354	9.383139
BF	1.7842	1.7492	0.0239	0.1573	0.8527	0.9689	5.2163
WBCID	0.029583	1.178055	0.0000	-0.3773	0.5656	**0.5791**	**0.6679**
GF	-3.0704	1.8864	0.0526	0.1330	0.7693	0.9714	2.8796
JBF	1.6775	1.6466	0.0123	0.1661	0.8813	0.9608	3.3174
Kim	3.6721	1.5126	0.0000	0.0145	0.8291	0.8417	1.6261
NHR	**14.2233**	**4.0722**	0.0142	0.1540	0.5631	0.9199	32.3782
He et al.	-0.7771	1.5615	0.0247	0.2261	0.8820	0.9835	20.4726
Tarel	12.9742	2.0086	**0.0000**	0.4122	0.8452	0.9618	4.4537

Appendix D. Evaluation Results Based on Dark Foggy Scene (Live)

Algorithm	e	r	Σ	HCC	SSIM	UQI	Time
Dehazenet	19.9871	1.3402	0.0128	-0.276955	0.658398	0.4827	3.2670
MSCNN	7.0346	1.1566	0.0008	0.539402	0.9625	0.938295	2.7783
Colores	14.1258	1.7970	0.0024	0.6412	0.8554	0.904098	2.1217
Zhu	16.705	1.2442	0.0004	-0.184528	0.7691	0.624096	2.3334
Multiband	23.0579	3.3220	0.0228	0.0120	0.5338	0.702196	1.8406
CODHWT	9.465	1.258896	0.001844	0.298458	0.883852	0.818569	1.971623
Meng	25.232083	3.178712	0.016137	-0.127894	0.577913	0.7351	4.841839
Liu	17.58625	1.945983	0.008957	-0.24152	0.603954	**0.555025**	0.852609
Berman	20.094583	3.06916	0.027333	-0.098533	**0.519443**	0.590679	7.147162
BF	10.8713	1.5772	0.0024	0.794161	0.916754	0.9769	5.2852
WBCID	3.6238	1.0175	0.0015	0.123286	0.846532	0.8484	**0.6304**
GF	9.8704	1.4801	0.0015	0.765783	0.930932	0.9742	3.8447
JBF	10.2158	1.5321	0.0018	**0.798623**	0.92159	0.9766	3.0011
Kim	15.129583	1.1581	**0.0000**	-0.081301	0.812525	0.7123	1.7037
NHR	26.8883	**4.8943**	0.0152	0.353259	0.480202	0.9024	33.3985
He et al.	11.6904	1.5489	0.0016	0.7082	0.915269	0.9625	18.6767
Tarel	**27.8892**	2.4675	**0.0000**	0.249464	0.779644	0.8962	4.4015

Appendix E. Evaluation Results Based on Sky Foggy Scene (Live)

Algorithm	e	r	Σ	HCC	SSIM	UQI	Time
Dehazenet	4.7113	0.9589	0.0043	0.307916	0.936388	0.9025	2.6442
MSCNN	5.7746	1.0628	0.0008	0.213925	0.9187	0.902131	1.9396
Colores	5.8375	1.1903	0.0003	-0.0806	0.9206	0.931895	2.6293
Zhu	4.080417	0.9243	**0.0000**	**0.334325**	0.9682	0.974045	2.2935
Multiband	7.3025	2.3827	0.0020	0.2544	0.7084	0.85643	0.8098
CODHWT	4.11625	0.840657	**0.0000**	0.315035	0.928537	0.891593	1.613
Meng	15.6625	1.801919	0.000208	0.080183	0.83313	0.884057	4.325442
Liu	6.89375	1.286005	0.00025	0.109946	0.779111	0.740609	0.857336
Berman	9.845417	1.771679	0.005214	0.017815	0.767244	0.843084	10.199034
BF	-10.2125	1.0394	0.2444	-0.054068	0.713872	0.8511	5.4393
WBCID	-6.2750	0.5114	**0.0000**	0.085668	0.680668	0.8851	**0.6714**
GF	-7.2796	1.7923	0.2456	-0.056397	**0.630068**	**0.8385**	3.3170
JBF	-5.0350	1.7363	0.2176	-0.056235	0.671786	0.8619	3.3787
Kim	3.1346	0.9974	**0.0000**	0.140081	0.936219	0.9598	1.6741
NHR	6.3183	**2.7691**	0.1241	-0.047596	0.688985	0.9293	25.5386
He et al.	4.3975	1.0602	0.0411	-0.038059	0.956283	0.9809	18.7453
Tarel	**11.0075**	1.7998	**0.0000**	0.115319	0.798529	0.8601	5.3438

Appendix F. Evaluation Results Based on Inhomogeneous Foggy Scene (RESIDE)

Algorithm	e	r	Σ	HCC	SSIM	UQI	Time
Dehazenet	7.935049	0.854451	0.0185	0.1359	0.7425	0.6672	2.3510
MSCNN	9.297229	1.0999	0.0079	0.3668	0.8215	0.7970	1.5016
Colores	7.416572	1.3486	0.0012	0.3577	0.8951	0.9400	2.4097
Zhu	1.9523	1.1116	0.0001	**0.5563**	0.8819	0.9077	1.5011
Multiband	22.276112	2.4266	0.0100	0.2261	**0.6016**	0.7249	0.8679
CODHWT	7.989725	0.863739	0.00167	0.22102	0.753898	0.695235	1.310947
Meng	15.409125	1.823188	0.02838	-0.109583	0.763196	0.794093	4.727146
Liu	13.722191	1.501098	0.003609	0.021543	0.65972	0.60865	0.866733
Berman	20.383201	1.976998	0.026848	-0.071128	0.511494	**0.558347**	3.451106
BF	10.460973	1.49539	0.0132	0.0101	0.8345	0.8679	5.1954
WBCID	-4.245852	0.605494	**0.0000**	0.2434	0.6810	0.8055	**0.5932**
GF	5.001414	1.158141	0.0203	-0.0101	0.8522	0.8806	4.0565
JBF	3.49736	1.015395	**0.0000**	0.3248	0.8468	0.8591	3.7495
Kim	-0.776772	1.249675	0.0002	0.4708	0.9155	0.9757	1.8187
NHR	**25.607089**	**4.095499**	0.0234	-0.1204	0.5311	0.8973	15.5492
He et al.	3.565705	1.258651	0.0032	0.2087	0.9067	0.9587	18.0359
Tarel	19.536199	3.357711	0.0002	0.4999	0.6243	0.8480	3.6023

Appendix G. Evaluation Results Based on Homogeneous Foggy Scene (RESIDE)

Algorithm	e	r	Σ	HCC	SSIM	UQI	Time
Dehazenet	3.5388	1.9113	**0.0000**	0.7258	0.9314	0.9787	2.3979
MSCNN	9.1789	2.8502	**0.0000**	0.814069	0.868812	0.9589	1.8033
Colores	12.5962	4.0872	0.0034	0.2654	0.718492	0.88795	2.38427
Zhu	2.6065	1.7850	**0.00000**	0.7580	0.941935	0.984282	2.212608
Multiband	13.9621	4.8493	**0.0000**	0.7529	0.738212	0.910496	0.848164
CODHWT	1.014329	1.418885	0	**0.939925**	0.971873	0.99807	1.2825
Meng	32.944476	8.405718	0.012264	-0.086677	**0.541719**	0.837048	3.809055
Liu	16.002074	5.159446	0.000105	0.398135	0.675243	0.908446	0.821688
Berman	24.737462	6.874965	0.00005	0.279624	0.576514	0.860761	3.001658
BF	7.4618	2.5616	**0.0000**	0.5586	0.8851	0.9829	4.6799
WBCID	3.3211	2.2371	**0.0000**	0.3524	0.8625	0.9354	**0.5807**
GF	5.7669	3.1820	0.2470	-0.0186	0.8835	0.9291	3.8348
JBF	6.6431	3.0093	0.1427	-0.0065	0.8715	0.9519	4.1577
Kim	9.2671	3.2265	**0.0000**	0.4194	0.8546	0.9577	1.4141
NHR	**47.6654**	**9.1384**	0.0004	-0.2394	0.3228	**0.4333**	34.7232
He et al.	5.6439	2.8043	0.1636	-0.0118	0.8850	0.9471	18.1905
Tarel	11.8085	3.7466	**0.0000**	0.0246	0.7963	0.9581	3.5908

Appendix H. Evaluation Results Based on Dark Foggy Scene (RESIDE)

Algorithm	e	r	Σ	HCC	SSIM	UQI	Time
Dehazenet	6.7713	1.0578	0.0160	0.404655	0.860955	0.7891	2.4317
MSCNN	6.7944	0.9424	0.0096	0.145465	**0.8040**	0.671941	1.7006
Colores	6.8755	1.1209	0.0004	0.5106	0.9387	0.869402	3.3024
Zhu	2.884615	0.9944	**0.0000**	**0.993043**	0.9801	0.985353	1.4978
Multiband	23.9621	**3.0297**	0.0009	0.6262	0.8405	0.914939	**0.6546**
CODHWT	2.631976	0.999548	0.00003	0.139021	0.934012	0.876721	1.307001
Meng	7.511312	1.671939	0.000184	0.71885	0.861989	0.88257	4.816429
Liu	4.564951	0.963581	0.000069	0.459224	0.950932	0.87996	0.813935
Berman	3.300339	1.168217	0.00231	0.65091	0.955972	0.932679	10.36607
BF	4.7125	1.0423	0.0001	0.848326	0.976514	0.9573	4.8856
WBCID	-1.2967	0.7573	**0.0000**	0.622145	0.968764	0.9672	0.6671
GF	3.4106	0.9641	0.0001	0.865405	0.976906	0.9589	3.1854
JBF	3.3965	0.9655	0.0001	0.865081	0.976817	0.9588	4.0713
Kim	5.5543	1.0219	0.0001	0.97595	0.969048	0.9259	1.4256
NHR	**32.2643**	1.5287	0.0034	0.509224	0.565946	**0.4016**	16.3898
He et al.	4.3670	1.0412	0.0005	0.960589	0.981552	0.9745	17.3223
Tarel	8.6732	1.3766	**0.0000**	0.333919	0.931369	0.9044	3.5156

Appendix I. Evaluation Results Based on Sky Foggy Scene (RESIDE)

Algorithm	e	r	Σ	HCC	SSIM	UQI	Time
Dehazenet	4.5692	0.9386	0.0048	0.246052	0.784905	0.7265	2.2933
MSCNN	9.1987	1.3668	**0.0000**	0.305573	0.8480	0.886879	1.6186
Colores	6.5031	1.2967	0.0002	0.0367	0.8445	0.847032	2.1885
Zhu	1.657711	0.9709	**0.0000**	0.293672	0.8594	0.87639	2.2368
Multiband	21.6893	3.1120	0.0006	-0.1611	0.5866	0.742517	0.8356
CODHWT	4.471625	1.032288	0.00154	0.358459	0.816675	0.786138	1.31625
Meng	9.467383	1.862447	0.029514	-0.16236	0.80443	0.884386	4.338396
Liu et al.	18.289498	1.617126	0.000814	-0.464062	0.547117	**0.489511**	0.818556
Berman	20.034408	2.400936	0.00219	-0.232462	0.614909	0.680839	9.437322
BF	-0.6820	1.3640	0.1588	-0.090832	0.85684	0.9283	5.5421
WBCID	8.5831	1.1935	0.0002	-0.241839	**0.486279**	0.6201	**0.6067**
GF	1.8010	1.4817	0.1355	-0.089848	0.858631	0.9431	3.1114
JBF	7.6574	4.0109	0.1737	-0.108979	0.638098	0.8393	3.2164
Kim	3.2589	1.2584	**0.0000**	**0.451376**	0.882397	0.9451	1.4026
NHR	**24.2605**	**4.8230**	0.0105	-0.080688	0.534496	0.9346	24.2869
He et al.	2.9044	1.1994	0.0505	-0.089424	0.888942	0.9503	17.5200
Tarel	19.7459	2.8175	**0.0000**	0.299531	0.695741	0.8827	4.2647

References

[1] S. D. Roy and M. K. Bhowmik, "A survey on visibility enhancement techniques in degraded atmospheric outdoor scenes," in 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017, pp. 349-352.

[2] S. D. Roy, M. K. Bhowmik, and S. S. Saha, "Qualitative evaluation of visibility enhancement techniques on SAMEER-TU database for security and surveillance," in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1-7.

[3] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," IEEE transactions on pattern analysis and machine intelligence, vol. 25, no. 6, pp. 713-724, 2003.

[4] Q. Zhu, J. Mai, and L. Shao, "A Fast Single Image Haze Removal Algorithm Using Color Attenuation Prior," IEEE Trans. Image Processing, vol. 24, no. 11, pp. 3522-3533, 2015.

[5] D. Nair and P. Sankaran, "Color image dehazing using surround filter and dark channel prior," Journal of Visual Communication and Image Representation, vol. 50, pp. 9-15, 2018.

[6] Y. Xu, J. Wen, L. Fei, and Z. Zhang, "Review of Video and Image Defogging Algorithms and Related Studies on Image Restoration and Enhancement," IEEE Access, vol. 4, pp. 165-188, 2016.

[7] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," International Journal of Computer Vision, vol. 48, no. 3, pp. 233-254, 2002.

[8] W. Wang and X. Yuan, "Recent advances in image dehazing," IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 3, pp. 410-436, 2017.

[9] A. K. Tripathi and S. Mukhopadhyay, "Single image fog removal using anisotropic diffusion," IET Image Processing, vol. 6, no. 7, pp. 966-975, 2012.

[10] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 12, pp. 2341-2353, 2011.

[11] R. Fattal, "Dehazing using color-lines," ACM transactions on graphics (TOG), vol. 34, no. 1, p. 13, 2014.

[12] Z. Lin and X. Wang, "Dehazing for image and video using guided filter," Open J. Appl. Sci, vol. 2, no. 4B, pp. 123-127, 2012.

[13] G. Yadav, S. Maheshwari, and A. Agarwal, "Fog removal techniques from images: A comparative review and future directions," in 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT 2014), 2014, pp. 44-52.

[14] K. B. Gibson, D. T. Vo, and T. Q. Nguyen, "An investigation of dehazing effects on image and video coding," IEEE transactions on image processing, vol. 21, no. 2, pp. 662-673, 2012.

[15] S.-C. Huang, B.-H. Chen, and W.-J. Wang, "Visibility restoration of single hazy images captured in real-world weather conditions," IEEE Transactions on Circuits and Systems for Video Technology, vol. 24, no. 10, pp. 1814-1824, 2014.

[16] M. Wang and S.-d. Zhou, "The study of color image defogging based on wavelet transform and single scale retinex," in International Symposium on Photoelectronic Detection and Imaging 2011: Advances in Imaging Detectors and Applications, 2011, vol. 8194, p. 81940F: International Society for Optics and Photonics.

[17] X. Zhao, R. Wang, and Y. Qiu, "An enhancement method of fog-degraded images," in Second International Conference on Digital Image Processing, 2010, vol. 7546, p. 75461S: International Society for Optics and Photonics.

[18] M.-Z. Zhu, B.-W. He, and L.-W. Zhang, "Atmospheric light estimation in hazy images based on color-plane model," Computer Vision and Image Understanding, vol. 165, pp. 33-42, 2017.

[19] P. Bekaert, T. Haber, C. Ancuti, and C. Ancuti, "Enhancing underwater images and videos by fusion," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 81-88: IEEE.

[20] X. Fu, Y. Huang, D. Zeng, X.-P. Zhang, and X. Ding, "A fusion-based enhancing approach for single sandstorm image," in Multimedia Signal Processing (MMSP), 2014 IEEE 16th International Workshop on, 2014, pp. 1-5: IEEE.

[21] R. Fattal, "Single image dehazing," ACM transactions on graphics (TOG), vol. 27, no. 3, p. 72, 2008.

[22] Q. Zhu, Z. Hu, and K. Ivanov, "Quantitative assessment mechanism transcending visual perceptual evaluation for image dehazing," in 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2015, pp. 808-813.

[23] Z. Y. Hu and Q. Liu, "A Method for Dehazed Image Quality Assessment," in Practical Applications of Intelligent Systems, Iske 2013, vol. 279, Z. Wen and T. Li, Eds. Advances in Intelligent Systems and Computing, 2014, pp. 909-913.

[24] J. Mai, Q. Zhu, and D. Wu, "The latest challenges and opportunities in the current single image dehazing algorithms," in 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014), 2014, pp. 118-123.

[25] B. Li et al., "Benchmarking Single-Image Dehazing and Beyond," IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 492-505, 2019.

[26] T. Pal, M. K. Bhowmik, D. Bhattacharjee, and A. K. Ghosh, "Visibility enhancement techniques for fog degraded images: A comparative analysis with performance evaluation," in 2016 IEEE Region 10 Conference (TENCON), 2016, pp. 2583-2588.

[27] J. Kaur and P. Kaur, "Comparative study on various single image defogging techniques," in 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017, pp. 357-361.

[28] K. H. Abdulkareem et al., "A new standardisation and selection framework for real-time image dehazing algorithms from multi-foggy scenes based on fuzzy Delphi and hybrid multi-criteria decision analysis methods," Neural Computing and Applications, 2020/05/26 2020.

[29] K. H. Abdulkareem, et al. , "A Novel Multi-Perspective Benchmarking Framework for Selecting Image Dehazing Intelligent Algorithms Based on BWM and Group VIKOR Techniques," International Journal of Information Technology & Decision Making, vol. Vol. 19, 2020.

[30] K. Wang, H. Wang, Y. Li, Y. Hu, and Y. Li, "Quantitative Performance Evaluation for Dehazing Algorithms on Synthetic Outdoor Hazy Images," IEEE Access, vol. 6, pp. 20481-20496, 2018.

[31] J. Perez, P. J. Sanz, M. Bryson, and S. B. Williams, "A benchmarking study on single image dehazing techniques for underwater autonomous vehicles," in OCEANS 2017 - Aberdeen, 2017, pp. 1-9.

[32] Z. A. Hu, Q. S. Zhu, and Ieee, "AN EFFECTIVE PERFORMANCE RANKING MECHANISM TO IMAGE DEHAZING METHODS WITH PSYCHOLOGICAL INFERENCE BENCHMARK," in 2016 Ieee International Conference on Acoustics, Speech and Signal Processing Proceedings(International Conference on Acoustics Speech and Signal Processing ICASSP, 2016, pp. 1576-1580.

[33] Z. Chen, T. Jiang, and Y. Tian, "Quality Assessment for Comparing Image Enhancement Algorithms," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3003-3010.

[34] B. Li, M. Tian, W. Zhang, H. Yao, X. J. J. o. V. C. Wang, and I. Representation, "Learning to predict the quality of distorted-then-compressed images via a deep neural network," vol. 76, p. 103004, 2021.

[35] X. Liu and J. Y. Hardeberg, "Fog removal algorithms: Survey and perceptual evaluation," in European Workshop on Visual Information Processing (EUVIP), 2013, pp. 118-123.

[36] K. D. Ma, W. T. Liu, Z. Wang, and Ieee, "PERCEPTUAL EVALUATION OF SINGLE IMAGE DEHAZING ALGORITHMS," in 2015 Ieee International Conference on Image Processing(IEEE International Conference on Image Processing ICIP, 2015, pp. 3600-3604.

[37] J. El Khoury, S. Le Moan, J.-B. Thomas, A. J. M. t. Mansouri, and applications, "Color and sharpness assessment of single image dehazing," vol. 77, no. 12, pp. 15409-15430, 2018.

[38] C. H. Hsieh, S. C. Horng, Z. J. Huang, and Q. Zhao, "Objective Haze Removal Assessment Based on Two-Objective Optimization," in 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), 2017, pp. 279-283.

[39] F. Guo, J. Tang, and Z. X. Cai, "Objective measurement for image defogging algorithms," Journal of Central South University, vol. 21, no. 1, pp. 272-286, Jan 2014.

[40] Y. Wang et al., "An imaging-inspired no-reference underwater color image quality assessment metric," Computers & Electrical Engineering, 2017.

[41] S. Fang, J. R. Yang, J. Q. Zhan, H. W. Yuan, R. Z. Rao, and Ieee, "Image Quality Assessment on Image Haze Removal," in 2011 Chinese Control and Decision Conference, pp. 610-614.

[42] X. X. Pan, F. Y. Xie, Z. G. Jiang, Z. W. Shi, and X. Y. Luo, "No-Reference Assessment on Haze for Remote-Sensing Images," Ieee Geoscience and

Remote Sensing Letters, vol. 13, no. 12, pp. 1855-1859, Dec 2016.

[43] K. Li, Y. Li, S. You, and N. Barnes, "Photo-Realistic Simulation of Road Scene for Data-Driven Methods in Bad Weather," in 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 491-500.

[44] J. El Khoury, J. B. Thomas, and A. Mansouri, "A Color Image Database for Haze Model and Dehazing Methods Evaluation," in Image and Signal Processing, vol. 9680, A. Mansouri, F. Nouboud, A. Chalifour, D. Mammass, J. Meunier, and A. ElMoataz, Eds. Lecture Notes in Computer Science, 2016, pp. 109-117.

[45] C. Ancuti, C. O. Ancuti, and C. D. Vleeschouwer, "D-HAZY: A dataset to evaluate quantitatively dehazing algorithms," in 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 2226-2230.

[46] A. Duarte, F. Codevilla, J. D. O. Gaya, and S. S. C. Botelho, "A dataset to evaluate underwater image restoration methods," in OCEANS 2016 - Shanghai, 2016, pp. 1-6.

[47] T. Pal, M. K. Bhowmik, and A. K. Ghosh, "Defogging of Visual Images Using SAMEER-TU Database," Procedia Computer Science, vol. 46, pp. 1676-1683, 2015.

[48] S. H. Wang, Y. Tian, T. Pu, P. Wang, and P. Perner, "A Hazy Image Database with Analysis of the Frequency Magnitude," International Journal of Pattern Recognition and Artificial Intelligence, vol. 32, no. 5, May 2018, Art. no. 1854012.

[49] Y. Li, K. Wang, N. Xu, and Y. Li, "Quantitative evaluation for dehazing algorithms on synthetic outdoor hazy dataset," in 2017 IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1-4.

[50] T. Zhang, H. M. Hu, and B. Li, "A Naturalness Preserved Fast Dehazing Algorithm Using HSV Color Space," IEEE Access, vol. PP, no. 99, pp. 1-1, 2018.

[51] H. Zhao, C. Xiao, J. Yu, and X. Xu, "Single image fog removal based on local extrema," IEEE/CAA Journal of Automatica Sinica, vol. 2, no. 2, pp. 158-165, 2015.

[52] S. B. Williams et al., "Monitoring of benthic reference sites: using an autonomous underwater vehicle," IEEE Robotics & Automation Magazine, vol. 19, no. 1, pp. 73-84, 2012.

[53] J.-P. Tarel, N. Hautiere, A. Cord, D. Gruyer, and H. Halmaoui, "Improved visibility of road scene images under heterogeneous fog," in 2010 IEEE Intelligent Vehicles Symposium, 2010, pp. 478-485: IEEE.

[54] J.-P. Tarel, N. Hautiere, L. Caraffa, A. Cord, H. Halmaoui, and D. J. I. I. T. S. M. Gruyer, "Vision enhancement in homogeneous and heterogeneous fog," vol. 4, no. 2, pp. 6-20, 2012.

[55] Y. Li, S. You, M. S. Brown, and R. T. Tan, "Haze visibility enhancement: A Survey and quantitative benchmarking," Computer Vision and Image Understanding, vol. 165, pp. 1-16, 2017.

[56] X. Zhu, Y. Li, and Y. Qiao, "Fast single image dehazing through Edge-Guided Interpolated Filter," in 2015 14th IAPR International Conference on Machine Vision Applications (MVA), 2015, pp. 443-446.

[57] T. M. Bui and W. Kim, "Single Image Dehazing Using Color Ellipsoid Prior," IEEE Transactions on Image Processing, vol. 27, no. 2, pp. 999-1009, 2018.

[58] S. Liu, M. A. Rahman, C. Y. Wong, S. C. F. Lin, G. Jiang, and N. Kwok, "Dark channel prior based image de-hazing: A review," in 2015 5th International Conference on Information Science and Technology (ICIST), 2015, pp. 345-350.

[59] X. Deng, H. Wang, X. Liu, and Q. Gu, "State of the art of the underwater image processing methods," in 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2017, pp. 1-6.

[60] M. Han, Z. Lyu, T. Qiu, and M. Xu, "A Review on Intelligence Dehazing and Color Restoration for Underwater Images," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. PP, no. 99, pp. 1-13, 2018.

[61] G. H. Babu, N. J. J. o. V. C. Venkatram, and I. Representation, "A survey on analysis and implementation of state-of-the-art haze removal techniques," p. 102912, 2020.

[62] A. C. Aponso and N. Krishnarajah, "Review on state of art image enhancement and restoration methods for a vision based driver assistance system with De-weathering," in 2011 International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2011, pp. 135-140.

[63] C. Chengtao, Z. Qiuyu, L. Yanhua, and Ieee, "A Survey of Image Dehazing Approaches," in 2015 27th Chinese Control and Decision Conference, 2015, pp. 3964-3969.

[64] D. Wu, Q. Zhu, J. Wang, Y. Xie, and L. Wang, "Image haze removal: Status, challenges and prospects," in 2014 4th IEEE International Conference on Information Science and Technology, 2014, pp. 492-497.

[65] S. Anwar and C. J. S. P. I. C. Li, "Diving deeper into underwater image enhancement: A survey," vol. 89, p. 115978, 2020.

[66] W. Rong and Y. XiaoGang, "A fast method of foggy image enhancement," in Proceedings of 2012 International Conference on Measurement, Information and Control, 2012, vol. 2, pp. 883-887.

[67] Y. H. Shiau, P. Y. Chen, H. Y. Yang, C. H. Chen, and S. S. Wang, "Weighted haze removal method with halo prevention," Journal of Visual Communication and Image Representation, vol. 25, no. 2, pp. 445-453, 2014.

[68] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An End-to-End System for Single Image Haze Removal," IEEE Transactions on Image Processing, vol. 25, no. 11, pp. 5187-5198, 2016.

[69] E. Zhang, K. Lv, Y. Li, and J. Duan, "A fast video image defogging algorithm based on dark channel prior," in 2013 6th International Congress on Image and Signal Processing (CISP), 2013, vol. 01, pp. 219-223.

[70] X. Liu, H. Zhang, Y. Y. Tang, and J. X. Du, "Scene-adaptive single image dehazing via opening dark channel model," IET Image Processing, vol. 10, no. 11, pp. 877-884, 2016.

[71] M. M. El-Hashash, H. A. Aly, T. A. Mahmoud, and W. Swelam, "A video haze removal system on heterogeneous cores," in 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2015, pp. 1255-1259.

[72] M. Negru, S. Nedevschi, and R. I. Peter, "Exponential Contrast Restoration in Fog Conditions for Driving Assistance," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 4, pp. 2257-2268, 2015.

[73] M. Negru, S. Nedevschi, and R. I. Peter, "Exponential image enhancement in daytime fog conditions," in 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2014, pp. 1675-1681.

[74] K. Roy, S. Kumar, S. Banerjee, T. S. Sarkar, and S. S. Chaudhuri, "Dehazing technique for natural scene image based on color analysis and restoration with road edge detection," in 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech), 2017, pp. 1-6.

[75] F. Guo, H. Peng, and J. Tang, "Fast Defogging and Restoration Assessment Approach to Road Scene Images," Journal of Information Science and Engineering, vol. 32, no. 3, pp. 677-702, May 2016.

[76] N. Hautiere, J. P. Tarel, and D. Aubert, "Mitigation of Visibility Loss for Advanced Camera-Based Driver Assistance," IEEE Transactions on Intelligent Transportation Systems, vol. 11, no. 2, pp. 474-484, 2010.

[77] W. Song, B. Deng, H. Zhang, Q. Xiao, and S. Peng, "An adaptive real-time video defogging method based on context-sensitiveness," in 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR), 2016, pp. 406-410.

[78] K. B. Gibson and T. Q. Nguyen, "An Analysis and Method for Contrast Enhancement Turbulence Mitigation," IEEE Transactions on Image Processing, vol. 23, no. 7, pp. 3179-3190, 2014.

[79] G. Yadav, S. Maheshwari, and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," in 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 2392-2397.

[80] B. Xie, F. Guo, and Z. X. Cai, "Universal strategy for surveillance video defogging," Optical Engineering, vol. 51, no. 10, Oct 2012, Art. no. 101703.

[81] B. Zhang and J. Zhao, "Hardware Implementation for Real-Time Haze Removal," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 25, no. 3, pp. 1188-1192, 2017.

[82] X. Zhao, W. Ding, C. Liu, and H. Li, "Haze removal for unmanned aerial vehicle aerial video based on spatial-temporal coherence optimisation," IET Image Processing, vol. 12, no. 1, pp. 88-97, 2018.

[83] W. Zhi, D. Watabe, and C. Jianting, "Improving visibility of a fast dehazing method," in 2016 World Automation Congress (WAC), 2016, pp. 1-6.

[84] H. Liu, D. Huang, S. Hou, and R. Yue, "Large size single image fast defogging and the real time video defogging FPGA architecture," Neurocomputing, vol. 269, pp. 97-107, 2017.

[85] Y. H. Shiau, Y. T. Kuo, P. Y. Chen, and F. Y. Hsu, "VLSI Design of an Efficient Flicker-Free Video Defogging Method for Real-Time Applications," IEEE Transactions on Circuits and Systems for Video Technology, vol. PP, no.

99, pp. 1-1, 2017.

[86] W. Wang, F. Chang, T. Ji, and X. Wu, "A Fast Single-Image Dehazing Method Based on a Physical Model and Gray Projection," IEEE Access, vol. 6, pp. 5641-5653, 2018.

[87] J. M. Guo, J. y. Syue, V. R. Radzicki, and H. Lee, "An Efficient Fusion-Based Defogging," IEEE Transactions on Image Processing, vol. 26, no. 9, pp. 4217-4228, 2017.

[88] B. Liao, P. Yin, and C. Xiao, "Efficient image dehazing using boundary conditions and local contrast," Computers & Graphics, vol. 70, pp. 242-250, 2018.

[89] Z. Gao and Y. Bai, "Single image haze removal algorithm using pixel-based airlight constraints," in 2016 22nd International Conference on Automation and Computing (ICAC), 2016, pp. 267-272.

[90] A. Kumari, H. Kodati, and S. K. Sahoo, "Fast and efficient contrast enhancement for real time video dehazing and defogging," in 2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI), 2015, pp. 1-5.

[91] B. Xie, F. Guo, and Z. Cai, "Improved Single Image Dehazing Using Dark Channel Prior and Multi-scale Retinex," in 2010 International Conference on Intelligent System Design and Engineering Application, 2010, vol. 1, pp. 848-851.

[92] S. Serikawa and H. Lu, "Underwater image dehazing using joint trilateral filter," Computers & Electrical Engineering, vol. 40, no. 1, pp. 41-50, 2014.

[93] L. Changli, F. Tanghuai, M. Xiao, Z. Zhen, W. Hongxin, and C. Lin, "An improved image defogging method based on dark channel prior," in 2017 2nd International Conference on Image, Vision and Computing (ICIVC), 2017, pp. 414-417.

[94] X. Qian and L. Han, "Fast image dehazing algorithm based on multiple filters," in 2014 10th International Conference on Natural Computation (ICNC), 2014, pp. 937-941.

[95] W. Sun, H. Wang, C. H. Sun, B. L. Guo, W. Y. Jia, and M. G. Sun, "Fast single image haze removal via local atmospheric light veil estimation," Computers & Electrical Engineering, vol. 46, pp. 371-383, Aug 2015.

[96] A. Alajarmeh, R. A. Salam, K. Abdulrahim, M. F. Marhusin, A. A. Zaidan, and B. B. Zaidan, "Real-time framework for image dehazing based on linear transmission and constant-time airlight estimation," Information Sciences, vol. 436–437, pp. 108-130, 2018.

[97] A. Kumari and S. K. Sahoo, "Fast single image and video deweathering using look-up-table approach," AEU - International Journal of Electronics and Communications, vol. 69, no. 12, pp. 1773-1782, 2015.

[98] C. O. Ancuti, C. Ancuti, and P. Bekaert, "Effective single image dehazing by fusion," in 2010 IEEE International Conference on Image Processing, 2010, pp. 3541-3544.

[99] C. O. Ancuti and C. Ancuti, "Single Image Dehazing by Multi-Scale Fusion," IEEE Transactions on Image Processing, vol. 22, no. 8, pp. 3271-3282, 2013.

[100] F. Guo, Z. Cai, B. Xie, and J. Tang, "Automatic Image Haze Removal Based on Luminance Component," in 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), 2010, pp. 1-4.

[101] X. Liu, F. Zeng, Z. Huang, and Y. Ji, "Single color image dehazing based on digital total variation filter with color transfer," in 2013 IEEE International Conference on Image Processing, 2013, pp. 909-913.

[102] J. Zhang, Y. Ding, Y. Yang, and J. Sun, "Real-time defog model based on visible and near-infrared information," in 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2016, pp. 1-6.

[103] I. Riaz, T. Yu, Y. Rehman, and H. Shin, "Single image dehazing via reliability guided fusion," Journal of Visual Communication and Image Representation, vol. 40, Part A, pp. 85-97, 2016.

[104] N. S. Pal, S. Lal, and K. Shinghal, "Visibility enhancement of images degraded by hazy weather conditions using modified non-local approach," Optik, vol. 163, pp. 99-113, 2018.

[105] D. Huang, K. Chen, J. Lu, and W. Wang, "Single Image Dehazing Based on Deep Neural Network," in 2017 International Conference on Computer Network, Electronic and Automation (ICCNEA), 2017, pp. 294-299.

[106] X. Jiang, J. Sun, H. Ding, and C. Li, "Video Image De-fogging Recognition Algorithm based on Recurrent Neural Network," IEEE Transactions on Industrial Informatics, vol. PP, no. 99, pp. 1-1, 2018.

[107] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in European conference on computer vision, 2016, pp. 154-169: Springer.

[108] S. Salazar-Colores, I. Cruz-Aceves, and J.-M. Ramos-Arreguin, "Single image dehazing using a multilayer perceptron," Journal of Electronic Imaging, vol. 27, no. 4, p. 043022, 2018.

[109] J. T. Kirk, Light and photosynthesis in aquatic ecosystems. Cambridge university press, 1994.

[110] J.-P. Tarel, N. Hautiere, L. Caraffa, A. Cord, H. Halmaoui, and D. Gruyer, "Vision enhancement in homogeneous and heterogeneous fog," IEEE Intelligent Transportation Systems Magazine, vol. 4, no. 2, pp. 6-20, 2012.

[111] H. Lu, Y. Li, Y. Zhang, M. Chen, S. Serikawa, and H. Kim, "Underwater optical image processing: a comprehensive review," Mobile networks and applications, vol. 22, no. 6, pp. 1204-1211, 2017.

[112] B. H. Chen, S. C. Huang, and F. C. Cheng, "A High-Efficiency and High-Speed Gain Intervention Refinement Filter for Haze Removal," Journal of Display Technology, vol. 12, no. 7, pp. 753-759, 2016.

[113] J.-H. Kim, W.-D. Jang, J.-Y. Sim, and C.-S. Kim, "Optimized contrast enhancement for real-time image and video dehazing," Journal of Visual Communication and Image Representation, vol. 24, no. 3, pp. 410-425, 2013.

[114] C.-H. Yeh, L.-W. Kang, M.-S. Lee, and C.-Y. Lin, "Haze effect removal from image via haze density estimation in optical model," Optics express, vol. 21, no. 22, pp. 27127-27141, 2013.

[115] Y.-H. Shiau, H.-Y. Yang, P.-Y. Chen, and Y.-Z. Chuang, "Hardware implementation of a fast and efficient haze removal method," IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 8, pp. 1369-1374, 2013.

[116] W. Wang, X. Yuan, X. Wu, and Y. Liu, "Fast Image Dehazing Method Based on Linear Transformation," IEEE Transactions on Multimedia, vol. 19, no. 6, pp. 1142-1155, 2017.

[117] K. B. Gibson and T. Q. Nguyen, "Fast single image fog removal using the adaptive Wiener filter," in 2013 IEEE International Conference on Image Processing, 2013, pp. 714-718.

[118] G. Ge, Z. Wei, and J. Zhao, "Fast single-image dehazing using linear transformation," Optik - International Journal for Light and Electron Optics, vol. 126, no. 21, pp. 3245-3252, 11// 2015.

[119] S. G. Narasimhan and S. K. Nayar, "Interactive (de) weathering of an image using physical models," in IEEE Workshop on color and photometric Methods in computer Vision, 2003, vol. 6, no. 6.4, p. 1: France.

[120] J. Kopf et al., Deep photo: Model-based photograph enhancement and viewing (no. 5). ACM, 2008.

[121] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Polarization-based vision through haze," Applied optics, vol. 42, no. 3, pp. 511-525, 2003.

[122] N. Sadhvi, A. Kumari, and T. A. Sudha, "Bi-orthogonal wavelet transform based single image visibility restoration on hazy scenes," in 2016 International Conference on Communication and Signal Processing (ICCSP), 2016, pp. 2199-2203.

[123] M. Wang, J. Mai, Y. Liang, R. Cai, T. Zhengjia, and Z. Zhang, "Component-Based Distributed Framework for Coherent and Real-Time Video Dehazing," in 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2017, vol. 1, pp. 321-324.

[124] Y. t. Liang, L. Li, K. b. Zhao, and J. h. Hu, "Defogging algorithm of color images based on Gaussian function weighted histogram specification," in 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), 2016, pp. 364-369.

[125] J. Mai, Q. Zhu, D. Wu, Y. Xie, and L. Wang, "Back propagation neural network dehazing," in 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014), 2014, pp. 1433-1438.

[126] S. Goswami, J. Kumar, and J. Goswami, "A hybrid approach for visibility enhancement in foggy image," in 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 175-180.

[127] J.-H. Yu and T.-J. Xiao, "Design and implementation of pipeline structure of image filtering process based on FPGA," Computer Engineering and Design, vol. 30, no. 18, pp. 4192-4194, 2009.

[128] T. Yu, I. Riaz, J. Piao, and H. Shin, "Real-time single image dehazing using block-to-pixel interpolation and adaptive dark channel prior," IET Image Processing, vol. 9, no. 9, pp. 725-734, 2015.

[129] A. G. Khodary, H. A. Aly, and Ieee, A New Image-Sequence Haze

Removal System Based on DM6446 Davinci Processor (2014 Ieee Global Conference on Signal and Information Processing). 2014, pp. 703-706.

[130] H. Koschmieder, "Theorie der horizontalen Sichtweite," Beitrage zur Physik der freien Atmosphare, pp. 33-53, 1924.

[131] W. Zhang, J. Liang, H. Ju, L. Ren, E. Qu, and Z. Wu, "A robust haze-removal scheme in polarimetric dehazing imaging based on automatic identification of sky region," Optics & Laser Technology, vol. 86, pp. 145-151, 2016.

[132] F. Jalled and I. Voronkov, "Object Detection Using Image Processing," arXiv preprint arXiv:1611.07791, 2016.

[133] G. De Novi, C. Melchiorri, J. Garcia, P. Sanz, P. Ridao, and G. Oliver, "A new approach for a reconfigurable autonomous underwater vehicle for intervention," in Systems conference, 2009 3rd annual IEEE, 2009, pp. 23-26: IEEE.

[134] S. Liu et al., "Image de-hazing from the perspective of noise filtering," Computers & Electrical Engineering, vol. 62, pp. 345-359, 2017.

[135] A. Shihavuddin, N. Gracias, R. Garcia, J. Escartin, and R. B. Pedersen, "Automated classification and thematic mapping of bacterial mats in the north sea," in OCEANS-Bergen, 2013 MTS/IEEE, 2013, pp. 1-8: IEEE.

[136] C. Balletti, C. Beltrame, E. Costa, F. Guerra, and P. Vernier, "UNDERWATER PHOTOGRAMMETRY AND 3D RECONSTRUCTION OF MARBLE CARGOS SHIPWRECK," International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 2015.

[137] N. Aliane, J. Fernandez, M. Mata, and S. Bemposta, "A system for traffic violation detection," Sensors, vol. 14, no. 11, pp. 22113-22127, 2014.

[138] A. H. Ashtari, M. J. Nordin, and M. Fathy, "An Iranian license plate recognition system based on color features," IEEE transactions on intelligent transportation systems, vol. 15, no. 4, pp. 1690-1705, 2014.

[139] S.-C. Huang, B.-H. Chen, and Y.-J. Cheng, "An efficient visibility enhancement algorithm for road scenes captured by intelligent transportation systems," IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 5, pp. 2321-2332, 2014.

[140] X. Jiang, H. Yao, S. Zhang, X. Lu, and W. Zeng, "Night video enhancement using improved dark channel prior," in ICIP, 2013, pp. 553-557.

[141] S. M. Shankaranarayana, K. Ram, A. Vinekar, K. Mitra, and M. Sivaprakasam, "Restoration of neonatal retinal images," 2016.

[142] W. Rui and W. Guoyu, "Medical X-ray image enhancement method based on dark channel prior," in Proceedings of the 5th International Conference on Bioinformatics and Computational Biology, 2017, pp. 38-41.

[143] S. Jeong and S. Lee, "The single image dehazing based on efficient transmission estimation," in 2013 IEEE International Conference on Consumer Electronics (ICCE), 2013, pp. 376-377.

[144] G. Woodell, D. J. Jobson, Z.-u. Rahman, and G. Hines, "Advanced image processing of aerial imagery," in Visual Information Processing XV, 2006, vol. 6246, p. 62460E: International Society for Optics and Photonics.

[145] Q. Zhu, Z. Song, Y. Xie, and L. Wang, "A novel recursive Bayesian learning-based method for the efficient and accurate segmentation of video with dynamic background," IEEE Trans. Image Processing, vol. 21, no. 9, pp. 3865-3876, 2012.

[146] W. Liu and D. Tao, "Multiview hessian regularization for image annotation," IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2676-2687, 2013.

[147] Q. Zhu, Z. Zhang, Z. Song, Y. Xie, and L. Wang, "A novel nonlinear regression approach for efficient and accurate image matting," IEEE Signal Processing Letters, vol. 20, no. 11, pp. 1078-1081, 2013.

[148] L. Tang and G. Shao, "Drone remote sensing for forestry research and practices," Journal of Forestry Research, vol. 26, no. 4, pp. 791-797, 2015.

[149] H. Lu et al., "Depth map reconstruction for underwater Kinect camera using inpainting and local image mode filtering," IEEE Access, vol. 5, pp. 7115-7122, 2017.

[150] C. Huang, D. Yang, R. Zhang, L. Wang, and L. Zhou, "Improved algorithm for image haze removal based on dark channel priority," Computers & Electrical Engineering, 2017.

[151] G. Woodell, D. J. Jobson, Z.-u. Rahman, and G. Hines, "Enhancement of imagery in poor visibility conditions," in Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense IV, 2005, vol. 5778, pp. 673-684: International Society for Optics and Photonics.

[152] X. Ji, Y. Feng, G. Liu, M. Dai, and C. Yin, "Real-Time Defogging Processing of Aerial Images," in 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), 2010, pp. 1-4.

[153] Y. Qiu and S. Wu, "Contrast-based stereoscopic images dehazing," in 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), 2015, pp. 597-602.

[154] X. Zhang, Z. Bu, H. Chen, and M. Liu, "Fast image dehazing using joint Local Linear sure-based filter and image fusion," in 2015 5th International Conference on Information Science and Technology (ICIST), 2015, pp. 192-197.

[155] W. Liu, F. Zhou, T. Lu, J. Duan, and G. Qiu, "Image Defogging Quality Assessment: Real-World Database and Method," IEEE Transactions on Image Processing, vol. 30, pp. 176-190, 2021.

[156] X. Min et al., "Quality Evaluation of Image Dehazing Methods Using Synthetic Hazy Images," IEEE Transactions on Multimedia, vol. 21, no. 9, pp. 2319-2333, 2019.

[157] C. O. Ancuti, A. Kis, and C. Ancuti, "Evaluation of image dehazing techniques based on a realistic benchmark," in 2019 International Symposium ELMAR, 2019, pp. 61-64.

[158] X. Min, G. Zhai, K. Gu, X. Yang, and X. Guan, "Objective Quality Evaluation of Dehazed Images," IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 8, pp. 2879-2892, 2019.

[159] C. O. Ancuti, C. Ancuti, and R. Timofte, "NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 444-445.

[160] P. Wang et al., "Task-driven Image Preprocessing Algorithm Evaluation Strategy," in 2020 7th International Conference on Dependable Systems and Their Applications (DSA), 2020, pp. 500-508.

[161] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-Haze: A Benchmark for Image Dehazing with Dense-Haze and Haze-Free Images," in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 1014-1018.

[162] P. Mahajan, V. Jakhetiya, P. Abrol, P. Lehana, B. N. Subudhi, and S. C. Guntuku, "Perceptual Quality Evaluation of Hazy Natural Images," IEEE Transactions on Industrial Informatics, pp. 1-1, 2021.

[163] N. A. Husain, M. S. M. Rahim, S. Kari, and H. Chaudhry, "VRHAZE: The Simulation of Synthetic Haze Based on Visibility Range for Dehazing Method in Single Image," in 2020 6th International Conference on Interactive Digital Media (ICIDM), 2020, pp. 1-7: IEEE.

[164] D. Berman, D. Levy, S. Avidan, T. Treibitz, "Underwater Single Image Color Restoration Using Haze-Lines and a New Quantitative Dataset," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 8, pp. 2822-2837, 2021.

[165] H. Li, J. Li, and W. Wang, "A fusion adversarial underwater image enhancement network with a public test dataset," 2019.

[166] C. Li et al., "An underwater image enhancement benchmark dataset and beyond," vol. 29, pp. 4376-4389, 2019.

[167] S. Zheng, J. Sun, Q. Liu, Y. Qi, and S. Zhang, "Overwater Image Dehazing via Cycle-Consistent Generative Adversarial Network," in Proceedings of the Asian Conference on Computer Vision, 2020.

[168] C. Ancuti, C. O. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images," in International Conference on Advanced Concepts for Intelligent Vision Systems, 2018, pp. 620-631: Springer.

[169] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: a dehazing benchmark with real hazy and haze-free outdoor images," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 754-762.

[170] C. Sakaridis, D. Dai, and L. J. I. J. o. C. V. Van Gool, "Semantic foggy scene understanding with synthetic data," vol. 126, no. 9, pp. 973-992, 2018.

[171] J. El Khoury, J.-B. Thomas, A. J. J. o. I. S. Mansouri, and Technology, "A database with reference for image dehazing evaluation," vol. 62, no. 1, pp. 10503-1-10503-13, 2018.

[172] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3888-3901, 2015.

[173] Y. Cho, J. Jeong, A. J. I. R. Kim, and A. Letters, "Model-assisted multiband fusion for single image enhancement and applications to robot vision," vol. 3, no. 4, pp. 2822-2829, 2018.

[174] J. He, C. Zhang, R. Yang, and K. Zhu, "Convex optimization for fast image

dehazing," in 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 2246-2250.

[175] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 617-624.

[176] X. Liu, H. Zhang, Y.-m. Cheung, X. You, and Y. Y. Tang, "Efficient single image dehazing and denoising: An efficient multi-scale correlated wavelet approach," Computer Vision and Image Understanding, vol. 162, pp. 23-33, 2017.

[177] D. Berman and S. Avidan, "Non-local image dehazing," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1674-1682.

[178] J. Yu, C. Xiao, and D. Li, "Physics-based fast single image fog removal," in IEEE 10th International Conference on Signal Processing Proceedings, 2010, pp. 1048-1052: IEEE.

[179] R. He, Z. Wang, H. Xiong, and D. D. Feng, "Single Image Dehazing with White Balance Correction and Image Decomposition," in 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA), 2012, pp. 1-7.

[180] K. He, J. Sun, X. J. I. t. o. p. a. Tang, and m. intelligence, "Guided image filtering," vol. 35, no. 6, pp. 1397-1409, 2012.

[181] C. Xiao and J. Gan, "Fast image dehazing using guided joint bilateral filter," The Visual Computer, vol. 28, no. 6-8, pp. 713-721, 2012.

[182] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 226-234.

[183] J.-P. Tarel and N. Hautiere, "Fast visibility restoration from a single color or gray level image," in Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 2201-2208: IEEE.

[184] A. Kumari and S. K. Sahoo, "Real Time Visibility Enhancement for Single Image Haze Removal," Procedia Computer Science, vol. 54, pp. 501-507, 2015.

[185] A. J. S. P. Galdran, "Image dehazing by artificial multiple-exposure image fusion," vol. 149, pp. 135-147, 2018.

[186] J. Zhang et al., "Image dehazing based on dark channel prior and brightness enhancement for agricultural remote sensing images from consumer-grade cameras," vol. 151, pp. 196-206, 2018.

[187] Y Gao, Y., Su, Y., Li, Q., & Li, J, "Single fog image restoration with multi-focus image fusion," vol. 55, pp. 586-595, 2018.

[188] Y. Zhu, G. Tang, X. Zhang, J. Jiang, and Q. J. N. Tian, "Haze removal method for natural restoration of images with sky," vol. 275, pp. 499-510, 2018.

[189] K. Kim et al., "Improvement of radiographic visibility using an image restoration method based on a simple radiographic scattering model for x-ray nondestructive testing," vol. 98, pp. 117-122, 2018.

[190] Zotin, A. G., "Fast algorithm of image enhancement based on multi-scale retinex," International Journal of Reasoning-based Intelligent Systems, vol. 12, no. 2, pp. 106-116, 2020.

[191] Q. Tang et al., "Nighttime image dehazing based on Retinex and dark channel prior using Taylor series expansion," vol. 202, p. 103086, 2021.

[192] T. Wang, L. Zhao, P. Huang, X. Zhang, and J. J. N. Xu, "Haze concentration adaptive network for image dehazing," vol. 439, pp. 75-85, 2021.

[193] Y. Liu, A. Wang, H. Zhou, and P. Jia, "Single nighttime image dehazing based on image decomposition," Signal Processing, vol. 183, p. 107986, 2021.

[194] B. Gui, Y. Zhu, and T. Zhen, "Adaptive single image dehazing method based on support vector machine," Journal of Visual Communication and Image Representation, vol. 70, p. 102792, 2020.

[195] Y. Chen, B. Song, X. Du, and N. Guizani, "The enhancement of catenary image with low visibility based on multi-feature fusion network in railway industry," Computer Communications, vol. 152, pp. 200-205, 2020.

[196] B. Qi, C. Yang, L. Tan, X. Luo, and F. Liu, "A novel haze image steganography method via cover-source switching," Journal of Visual Communication and Image Representation, vol. 70, p. 102814, 2020.

[197] Y. Wu, Y. Qin, Z. Wang, X. Ma, and Z. Cao, "Densely pyramidal residual network for UAV-based railway images dehazing," Neurocomputing, vol. 371, pp. 124-136, 2020.

[198] K. Metwaly, X. Li, T. Guo, and V. Monga, "NonLocal Channel Attention for NonHomogeneous Image Dehazing," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 1842-1851.

[199] A. Kumari, S. K. Sahoo, and M. C. Chinnaiah, "Fast and Efficient Visibility Restoration Technique for Single Image Dehazing and Defogging," IEEE Access, vol. 9, pp. 48131-48146, 2021.

[200] X. Zhao, T. Zhang, W. Chen, and W. Wu, "Image Dehazing Based on Haze Degree Classification," in 2020 Chinese Automation Congress (CAC), 2020, pp. 4186-4191.

[201] R. Chen, Y. Sheng, S. Wei, and D. Tang, "Research on Safe Distance Measuring Method of Front Vehicle in Foggy Environment," in 2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4), 2019, pp. 333-338.

[202] J. Zhang, Z. Lu, and M. Li, "Active Contour-Based Method for Finger-Vein Image Segmentation," IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 11, pp. 8656-8665, 2020.

[203] A. Mehra, M. Mandal, P. Narang, and V. Chamola, "ReViewNet: A Fast and Resource Optimized Network for Enabling Safe Autonomous Driving in Hazy Weather Conditions," IEEE Transactions on Intelligent Transportation Systems, pp. 1-11, 2020.

[204] L.-P. Yao, Z.-l. J. M. T. Pan, and Applications, "The Retinex-based image dehazing using a particle swarm optimization method," pp. 1-18, 2020.

[205] I. U. Afridi, T. Bashir, H. A. Khattak, T. M. Khan, and M. Imran, "Degraded image enhancement by image dehazing and Directional Filter Banks using Depth Image based Rendering for future free-view 3D-TV," PLOS ONE, vol. 14, no. 5, p. e0217246, 2019.

[206] X. Wang, C. Yang, J. Zhang, H. J. I. J. o. A. Song, and B. Engineering, "Image dehazing based on dark channel prior and brightness enhancement for agricultural monitoring," vol. 11, no. 2, pp. 170-176, 2018.

[207] Y. Guo, J. Chen, X. Ren, A. Wang, and W. J. I. T. o. I. P. Wang, "Joint Raindrop and Haze Removal From a Single Image," vol. 29, pp. 9508-9519, 2020.

Karrar Hameed Abdulkareem

Karrar Hameed Abdulkareem received the B.S. degree in computer science (Artificial Intelligence) from the University of Technology, Iraq, in 2007, and the M.S. degree in computer science (Internetworking Technology) from the Universiti Teknikal Malaysia Melaka (UTeM), Malaysia, in 2016. He Obtained Ph.D. degree in Information Technology from Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia. He has produced more than 44 articles into different ISI Web of Science journals, such as IEEE Internet of Things, Journal of King Saud University - Computer and Information Sciences, Computer methods and programs in biomedicine (Elsevier), Neural Computing and Applications (Springer), IEEE Access, Journal of infection and public health (Elsevier), International Journal of Information Technology & Decision Making (World Scientific), Computers, Materials & Continua (Tech Science Press), Soft Computing (Springer), and Sensors (MDPI). His research area includes Multi-Criteria Decision Making, Artificial Intelligence, Data Science, Fog Computing, and Cyber Security.

Nureize Arbaiy

Nureize Arbaiy she is currently with Software Engineering Department at the Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia (UTHM). She received her Dr. of Engineering from the Graduate School of Information, Production, and Systems at Waseda University, Japan. She received her B.S degree from the University of Technology Malaysia in 2001 and M.S degree from University Utara of Malaysia in 2004. Her research interests are on multi-criteria decision making, fuzzy logic, expert system, fuzzy regression analysis, and possibilistic theory.

Zainab Hussein Arif

Zainab Hussein Arif received the B.S. degree in Computer Science from the University of Al-Qadisiyah, Iraq, in 2016, and the M.S. degree in Information Technology from the Universiti Tenaga Nasional, (UNITEN), Malaysia in 2021. Her research interests are on Artificial Intelligence, Deep Learning, and Data Science.

Mohammed Nasser Al-Mhiqani

Al-Mhiqani M.N received his BSc. In Computer Science (Computer Networking) in 2014, and MSc in Computer Science from Universiti Teknikal Malaysia Melaka (UTeM) in 2015. Currently he is a Ph.D. candidate at UTeM. His research interests Cyber Security, Cyber Physical System Security, and Insider Threats in Cyber Security, Machine Learning, Deep learning, and image processing.

Mazin Abed Mohammed

Mazin Abed Mohammed received the B.Sc. degree in computer science from the University of Anbar, Iraq, in 2008, the M.Sc. degree in information technology from UNITEN, Malaysia, in 2011, and the Ph.D. degree in information technology from UTeM, Malaysia, in 2019. He is currently a Lecturer with the College of Computer Science and Information Technology, University of Anbar, Iraq. His research interests include artificial intelligence, biomedical computing, and optimization.

Seifedine Kadry

Seifedine Kadry (Senior Member, IEEE) received the bachelor’s degree from Lebanese University, in 1999, the dual M.S. degree from Reims University, France, in 2002, and EPFL, Lausanne, the Ph.D. degree from Blaise Pascal University, France, in 2007, and the H.D.R. degree from Rouen University, in 2017. His current research interests include data science, education using technology, system prognostics, stochastic systems, and probability and reliability analysis. He is an ABET Program Evaluator of computing and an ABET Program Evaluator of Engineering Tech.

Zaid Abdi Alkareem Alyasseri

Zaid Abdi Alkareem Alyasseri received the B.Sc. degree in computer science from Babylon University, in 2007, and the M.Sc. degree in computer science from University Science Malaysia (USM), in 2013. In 2019 he got Ph.D. degree in the field of artificial intelligence (brain-inspired computing) with the Computational Intelligence Research Group, School of Computer Sciences. He is also a Senior Lecturer with the University of Kufa, Iraq. His research interests are optimization, pattern recognition, EEG, brain–computer interface, signal and image processing, machine learning.

Foundations for the Design of a Creative System Based on the Analysis of the Main Techniques that Stimulate Human Creativity

L. De Garrido[1], J. J. Gómez Sanz[2], J. Pavón[2] *

[1] Architecture Department, and Psychobiology Department. Universitat de València (Spain)
[2] Institute of Knowledge Technology. Universidad Complutense de Madrid (Spain)

Received 23 September 2020 | Accepted 13 February 2021 | Published 3 March 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

This work presents the design of a computational system with creative capacity, based on the synthesis of the main methods that stimulate human creativity. When analyzing each method, a set of characteristics that the computer system must have in order to emulate a creative capacity has been suggested. In this way, by integrating all the suggestions in a structured way, it is possible to design the general architecture and functioning strategy of a computer system that has the incremental creative capacity of well-known creative methods. This computational system is designed as a multi-agent system, made up of two groups of agents, the *problem solving group* and the *creative group*, the first one exploring and evaluating paths for suitable solutions, the second implementing creative methods to generate new paths that are provided to the first group.

Keywords

Artificial Intelligence Creative System, Multi-agent System, Creativity, Lateral Thinking, Methods to Stimulate Human Creativity.

DOI: 10.9781/ijimai.2021.03.001

I. Introduction

The present work aims to explain the basis for a creative computational system based on the synthesis of well-known effective methods that promote human creative capacity.

In a previous work, the implementation of a multi-agent system (MAS) has been shown, capable of emulating the creative capacity of a brainstorming method [1]. Based on this experience, an additional step has been taken to design a computational model that integrates several methods that stimulate the human creative capacity. These methods have been analyzed in order to know the reasons why they stimulate creativity, and to identify a set of hints for the implementation of a creative computational system.

One of the first systematic reviews of methods for producing artificial creativity [2] classified them in three types: combination of familiar ideas (combinational creativity), taking a thinking style and tweaking it (exploratory creativity), and changing dimensions of an existing idea (transformational creativity). Boden [3] also distinguishes between the historical creativity (H), which produces new ideas not known to have been reported at all, and psychological/personal creativity (P), which produces new ideas to the person. Creativity can also be seen as a social construction [4]. It cannot be reduced to some formal properties, but the debate can be useful to enumerate conditions under which an external observer is more likely to consider a system as creative. This criterion is similar to one of the stances when discussing what is artificial intelligence (AI), “intelligence is in the eye of the observer” [5], meaning that a system is as intelligent as an external observer considers it to be.

The four Ps [6] approach is cited frequently to analyze creativity research relating to humans. These four Ps stand for Person (what makes the agent a creative one), Process (what actions need to be undertaken to be creative), Product (what kind of creation is expected), and Press (what cultural context is applied to determine something as creative or not). Computational creativity can be addressed from these four Ps for evaluation, but existing research does not always consider them. For instance, 75% of the papers in the 2014 International Conference of Computational Creativity did not make any reference to social or interactive aspects of creativity, and were more focused on Product and Process aspects [6] [7].

A review on the literature about evaluation of creativity shows arising debates about the definition of terms, lack of autonomy in existing systems, the cultural specificity of many judgments, and the potentially domain-specific nature of creativity [8]. While operational tests (e.g. statistical analysis of the product) have been used for this purpose, questionnaires are a more frequent evaluation tool. Following recommendations from [8], one can address each P with some rules of thumb that include criteria to choose each P depending on the kind of impact one wants to achieve.

Another theoretical framework [9], intends to characterize the different creative systems and concepts, such as uninspiration (failing to be creative in a valued way) and aberration (deviation from the norms) in order to support formal reasoning about creativity, and uses these two examples as illustrative ones. The work presented in this paper is more extensive in the account of techniques. As such, it is more useful as guidelines for those willing to get acquaintance

* Corresponding author.

E-mail addresses: info@luisdegarrido.com (L. de Garrido), jjgomez@ucm.es (J. J. Gómez Sanz), jpavon@ucm.es (J. Pavón).

for creativity techniques. Also, rather than formalizing, it aims to serve as inspiration for enhancing other works. Formal models are more precise, but they are harder to apply than the generic guidelines shown in this paper.

The main contribution of this paper is the definition of an architectural framework, as a MAS, to explore how each method contributes to the generation of creative solutions in an integrated way. This is achieved by first identifying some suggestions from the analysis of each method, which are used as the basis for the implementation of the creative computational system. Their integration is made by defining an architecture of a MAS, which is structured in two groups of agents, the "problem solving group" and the "creative group". This organization is inspired from the functioning of the brain, where the problem solving group would correspond to the executive control network (ECN), and the creative group to the working of the default mode network (DMN). The basic idea is that the first explores and evaluates a set of paths to find a solution. When these are not successful, control is given to the creative group, which will generate new paths for the problem solving group.

First of all, all known systems that stimulate human creativity have been compiled. Many of these systems are variations of others, so the most representative and effective have been chosen.

The following creative methods are analyzed and applied in the design of the agents:

- *Establish analogies with known problems*
- *Creativity matrix*
- *Problem solving*
- *Brainstorming*
- *Variants of Brainstorming*
- *Graphic Brainstorming*
- *www-Brainstorming*
- *Lateral thinking*
- *Parallel thinking*

Sections II to IX describe the analysis of these methods in order to identify elements and functionality for the implementation of the creative computational system. These pieces are organized and integrated as a MAS, which is presented afterwards in section X, as well as considerations for the representation of the information that is managed by the agents. This model is discussed in section XI, taking into consideration other relevant works in the area of computational creativity. The paper concludes with a discussion (section XII) and some final remarks in section XIII.

II. Guidelines for the Design of a Creative Computational System, Based on the Analysis of the Establishment of Analogies

The most common way to solve problems creatively is to identify analogies with other problems that have been previously solved [10].

From birth, human beings begin to interact with their environment, in order to learn to function in it. The first learning tool they develop is imitation and establishing analogies. Human beings begin to imitate other humans in their close environment in order to join the group as soon as possible. This basic mechanism has been enhanced by the human evolutionary system to such an extent that a large number of "mirror neurons" have been created in our brain, whose task is to imitate the activity of those around us. This basic mechanism helps us to survive, integrating ourselves into the group, imitating patterns of action that have apparently been successful for others. Therefore, imitating the actions of others allows us to solve problems in the same way that others have solved. This mechanism can therefore seem very uncreative; however, when applied at various levels of abstraction the results of applying analogies can be surprising and enormously creative.

Human beings, before creating new things, must begin to know the existing ones. The imitation process is very important and occurs just when the human brain is developing, especially in the first four years of our life. Therefore, the imitation mechanism is actually a mechanism that limits human creativity at an early stage and with a low level of abstraction. However, with the passage of time, humans begin to accumulate more and more experiences, and many of them are completely new and unknown to them, and also, due to the very fact of being unknown, they have not been able to learn by imitating the acts of other humans when faced with such novel events. However, humans are able to draw basic parallels between new facts and known facts, in the hope that previously acquired cognitive strategies may help.

It is clear that several types of analogies can be established, with different levels of abstraction. Some analogies have a low level of abstraction and are therefore very close (for example, the analogy between driving a car and driving a truck). Others have a medium level of abstraction and are not so close (for example, the analogy between the activity of a coroner and the activity of a police officer). On the other hand, other analogies can have a high level of abstraction, and can seem very distant, and even belong to different fields of knowledge (for example, the analogy between roasting a piece of meat so that it is tasty and digestible, and superficially burning a beam wood to protect it against fire).

In order to establish analogies in different degrees of abstraction, an adequate representation of knowledge must be previously made, in such a way that the same object must have a huge number of possible attributes, some of which may seem obvious or very general. These attributes must be classified at different levels of abstraction, and they must be able to be activated and deactivated depending on the level of reasoning desired when establishing analogies.

In addition, a learning system must be available. This system will enrich the possibilities, and in the same way, the ability to establish analogies.

As an example to illustrate this mechanism, consider the act of "grilling" food. Hundreds of descriptions can be given to the act of "grilling", and the more descriptions, the more possibilities to establish analogies with similar actions, or in different settings. Grilling can be defined as the act of exposing a food to contact with a very hot radiant surface (such as a frying pan or a griddle), or simply to thermal radiation of a certain intensity, in order to make it more digestible, or with a more attractive flavor for the palate. By defining the act of grilling in this way, analogies can be drawn with any food, as grilling will make it more digestible and tastier. In this way, humans have learned to roast any type of food, or even food leftovers, in order to make them more digestible and tasty. In this way a fish, meat, vegetable, etc. could be roasted.

In this way, human beings have learned to grill any food, and have found that grilling improves its digestion and flavor. However, the concept of grilling is much broader, and has more connotations, and therefore greater applicability.

Alternatively, the act of "grilling" could be defined as the act of exposing food (or any other organic element) to contact with a very hot radiant surface (such as a frying pan or griddle), or simply to thermal radiation of a certain intensity, in order to raise the temperature of the surface of a food to change its physical structure, and make it more resistant to thermal radiation, and thus protect the inside of the food from thermal action. In this way, when roasting a piece of meat its surface is altered, it becomes carbonated, and in doing so the interior of the meat is protected from the thermal action. As a consequence, the carbonated surface of the piece of meat is stiffer, more crispy and attractive to the palate, and its interior is protected from thermal action, thus preserving its nutritional structure. In other words, when roasting a piece of meat, its outer carbonated structure protects the inside, keeping its nutrients intact, and also the result is more digestible and tasty.

In this way humans have established direct parallels, and have learned to grill any type of food. But they have also learned to establish indirect parallels, and they have learned to roast any other type of organic elements, in order to alter their external structure to preserve their internal structure. For example, we know that if the surface portion of a wooden beam is burned, the wood burned around the edges hardens, helping to protect the interior of the wooden beam from fire. Therefore, learning to grill, we can learn to protect a beam against fire, creating analogies with a high level of abstraction. In fact, this action is called "heat-treated wood", and consists of oversizing the section of wood so that when there is a fire, only the perimeter bark is altered and this scorched bark protects the interior of the beam from the flames of the fire.

Basically this is what case based reasoning (CBR) systems do, although they should be adapted to work with various levels of abstraction. The algorithms that are used in a CBR system for retrieval of past cases, and their reuse and adaptation to the new problem are appropriate for the implementation of the establishment of analogies, which can take advantage of past experiences.

III. Guidelines for the Design of a Creative Computational System, Based on the Analysis of Creativity Matrices

The *creativity matrix* is a method of assured effectiveness to stimulate creativity in solving problems that have a limited range of possible solutions [11][12]. The method consists of a combinatorial explosion between the different possibilities of each aspect that we want to take into account in solving a certain problem. For example, a lighter can be designed with only two attributes in mind: the fuel and the ignition system. By testing all possible values of each attribute and combining them with each other, we can get new and unexpected ideas.

From the analysis of the creativity matrices, it can be deduced that an important component of creative problem solving is the exploration of new search paths, never traveled before. These search paths are based on the association of certain characteristics of the fundamental parameters of a certain problem.

A first computational approach would consist in creating all possible combinations of all the relevant parameters of the objects to be designed. If the number of parameters is under a computational system, it could go through all its possible combinations and associations in a short moment of time, and choose the most appropriate combination, capable of providing the most creative solution. However, if the number of relevant parameters is very high, the possible combinations would grow explosively, the system would be very slow, and it would exhaust the patience of the possible users of the system. That is to say, what is usually called a combinatorial explosion would be created.

To avoid the combinatorial explosion, the different parameters must be evaluated in several different ways, in order to reduce the possible combinations, and make the use of the resulting computational system viable in the application of this system. The restrictions can be the following:

- Removing unwanted combinations of some parameters.
- Enhancing certain desired combinations.
- Assessing each parameter, to encourage, more or less, its use.
- Assessing combinations of combinations.
- Rating each parameter based on the preferences of each user.

In this way the possible combinations between parameters are considerably reduced, and therefore the method would be viable.

As a consequence, a first computational approach to this system would be the Heuristic Search and Solution Tree Pruning systems, although with certain important nuances.

Heuristic search in artificial intelligence is a technique for solving problems whose solution consists of a series of steps that frequently must be determined by systematic testing of alternatives. Therefore, it can be said that heuristic search algorithms are a computational method to solve path-finding problems, that is, "search for the best route from point A to point B" (see Fig. 1).

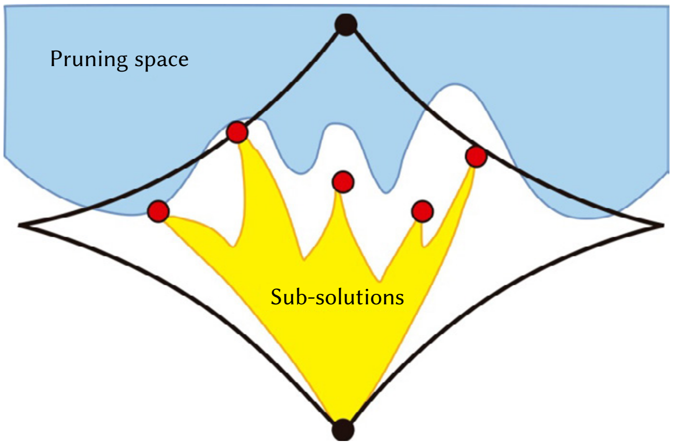


Fig. 1. Delimitation of conventional search areas using tree pruning of possible solutions.

Subsequently, the existence of an e*valuation function* that should measure the estimated distance to the target is assumed. This evaluation function is used to guide the process by selecting the most promising status or operations at each moment. This system does not always guarantee to find a solution, and if it is found it does not guarantee that it is the best.

Heuristic search methods have some information about the proximity of each state to a target state, allowing the most promising paths to be explored first. However, the conceptual functioning of a creativity matrix differs in some determining concepts. One of them is pure chance. Since each tour of the creativity matrix, presupposes that there are always valid solutions, depending on the specific combination of specific parameters that have been initially chosen. In other words, initially a specific combination of parameters is chosen by pure chance, or by a specific personal preference, or out of curiosity, and then a search process is initiated. We should keep in mind that exploring random paths, in the search for possible creative solutions, is a basic and recurring component of any method of stimulating creativity.

On the other hand, chance is also an essential component of the functioning of the human cognitive system. In most cases we are not aware that the brain explores more paths than we consciously explore, and many of them are explored at random (stimulated by concrete experiences during the creative process). In fact, when the brain does not have external stimuli that induce action and decision making, it simply has a default activity, in which it fiddles with possible scenarios determined many times by pure chance.

Therefore, from the analysis of the creativity matrices, two fundamental guidelines emerge in order to design a conceptual model of a creative computational system:

- The "*problem solving group*" must incorporate heuristic search mechanisms to avoid combinatorial explosion.
- The "*creative group*" must incorporate a module for generating random associations. These random associations can be of several types:
  - Absolute chance.
  - Chance limited by similarity (looking for combinations of components that have certain specific attributes in common).
  - Chance limited by strangeness (looking for combinations of components that apparently have nothing in common).
  - Chance limited by personal preferences (looking for combinations of components with certain attributes).

IV. Guidelines for the Design of a Creative Computational System, Based on the Analysis of Problem Solving Strategy (Vertical Thinking)

The problem-solving method is a purely deductive procedure based on cause-effect chains [13]. This thinking can be top-down, or bottom-up. In the first case, a problem is divided into sub-problems and an attempt is made to solve each sub-problem, often dividing it into more sub-problems, until finally all the partial solutions are concatenated, obtaining a possible final solution. In the second case, simple sub-problems are solved, joining with other sub-problems, until a certain situation is obtained that may coincide with an initial statement of the problem.

Vertical thinking, based on a problem solving structure, implies going through a certain path within the tree of possible solutions, from the stride to a certain leaf. Therefore, there is a huge number of possible paths within the decision tree until a possible decision is reached (see Fig. 2 and Fig. 3). This is why strategies must be established based on a pruning of the decision tree, choosing priorities to travel as a priority only certain paths until reaching a possible solution [14]. It is the most widely used approach in computer systems; in fact, it has a certain coincidence with the operation of rule-based systems.

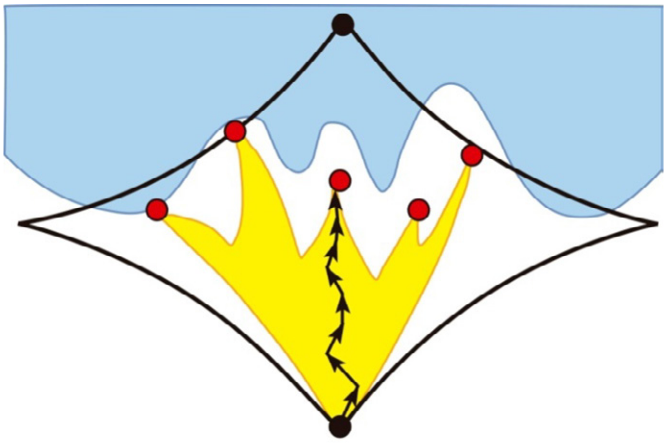


Fig. 2. Ideal and successful path to find solutions through conventional problem-solving systems.

However, the problem with rule-based systems is that they produce a multitude of solutions when the problem is complex and poorly defined. On the other hand, if the field of action is reduced so that it is well defined, the usual solutions are not creative at all.

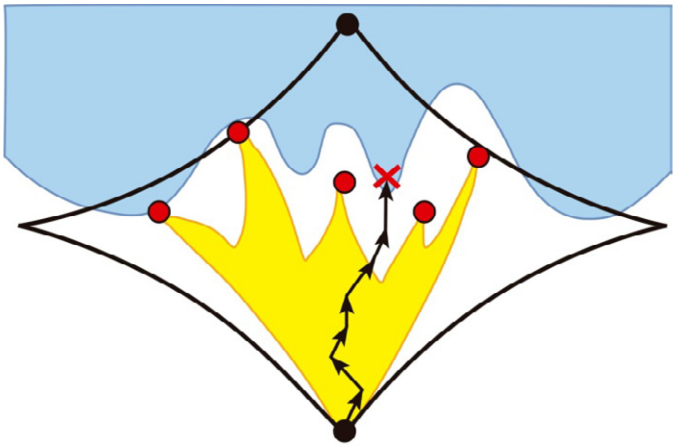


Fig. 3. Possible and unsuccessful paths to find solutions through conventional problem-solving systems.

In some cases, when the algorithms are very complex, the system could deduce some solution that might seem surprising or creative, simply due to the complex algorithmic interactions. However, creative solutions would be a rarity, since the system is based on the logical concatenation of the most appropriate actions in each case, which is why they tend to leave out a huge number of novel, surprising and therefore creative solutions.

For this reason, these problem solving systems must be complemented with other systems, which will be activated when it is not able to find novel solutions. In other words, systems based on the problem solving structure could come up with some creative solutions on their own, but in the event that it fails to generate sufficiently creative solutions, or simply generates an insufficient number, the system must give up control to other alternative systems.

The most important suggestion here is that the agents of the "*problem solving group*" should be implemented in a conventional way, following some rule-based technique, for instance. If the system can find a solution, but if it is not considered creative enough, this group cedes control to the "*creative group*", which has other more creative complementary structures.

V. Guidelines for the Design of a Creative Computational System, Based on the Analysis of Brainstorming

Brainstorming is probably the oldest and best known creative problem solving technique [15][16]. The main objective of *Brainstorming* is to break the usual limitations of the human cognitive system, and generate a large number of ideas, many of which can be really creative and solve a certain problem.

The structure of Brainstorming suggests that a MAS is a suitable computational model of a group of individuals that collaborate to find a solution. In this system, each agent does not have to be specialized in a specific task, but instead each agent must have a different structure and therefore different behavior. In other words, each agent must have a different algorithmic structure than the others, and therefore must process information differently. All agents share the information in a common space, in such a way that the information generated by a certain agent can be useful to another agent, who uses it and generates new information in the common space.

Based on the information generated at each moment, each agent can decide to work in a linear way, analyzing the information and deducing new information, or randomly, to a greater or lesser degree.

To design the basic architecture of the MAS, the functional structure of a Brainstorming must be analyzed, which is based on four fundamental rules.

A. Eliminate Criticism

The elimination of criticism in the idea generation process is an important factor to take into account when designing a creative computer system. This factor clearly indicates that to promote creative thinking, each possible solution should not be evaluated as soon as it is generated, since, even if it is a bad solution, it could serve as starting information for another agent who may be able to propose a creative solution. That is, ideas should not be eliminated early as they arise, since even if a given idea is not valid, it can stimulate any of the other agents to generate new ideas, which could be very creative.

B. Absolute Freedom

Absolute Freedom suggests several things.

- Each agent may have a different structure.
- The generation of ideas, although they may seem absurd, should be stimulated, since during the process, they could stimulate the generation of truly creative ideas that solve the problem. At any intermediate stage of the process, it might seem that the solutions that are being generated will not be adequate (if they were evaluated at the same time), but as the process continues, there could be unexpected changes and complex feedback that could lead to one or more valid unexpected solutions, new and therefore creative. For this reason, there should be some agent in charge of generating ideas randomly ("random agent").
- The ideas generated must be stored in a certain file, in order to be evaluated in a second stage.

C. Generate a Lot of Ideas

The computer system should create as many ideas as possible, in the common space of the MAS, in order to have a better chance that some of them can be successful. In addition, in the evaluation stage, the degree of creativity of each solution can be assessed, and in this way the system can learn, in order to explore paths close to those by which ideas have been generated considered as more creative. Therefore, the system should have a *learning system* that generates information for a certain agent, so that their subsequent proposals are based on previous successes.

D. Multiplying Effect

The ideas proposed by each agent throughout the process may have been generated as an extension of the ideas that they have previously proposed, but especially they must be provoked by the ideas proposed by the rest of the agents. That is, each agent develops its activity as a consequence of the activity previously developed by other agents. In this way, the ideas generated by each agent feed the rest of the agents, generating a multiplier effect of the ideas, which vary subtly, depending on the essential parameters of each agent.

It would therefore be advisable to have some agent generates ideas contrary to those generated by the other agents (which is not an evaluation or criticism of the rest of the ideas, it is simply a new idea contrary to those that are being proposed). This agent could be called "*tenth agent*" (the "tenth agent can be the same agent as the "random agent", but assuming a different role, therefore, from now on it will be called "*spark agent*"). Therefore, and as a result of this collaboration of agents, the system may be able to solve the problem in surprising ways.

Once the basic parameters of a MAS based on the creative structure of a *Brainstorming* have been established, it is convenient to analyze the development of a work session, and therefore determine the roles that each agent should have based on their operation.

1. Secretary

An agent must take care of the logistics of operation of the group of agents, organizing the order of intervention of each one, as well as their priorities ("*secretary agent*"). In the same way, it must be in charge of storing all the ideas generated by the set of agents.

2. Relaxed and Cheerful Atmosphere

Today it is known that a relaxed and joyful environment induces the brains of well-trained specialists to switch modes of operation, to deactivate the ECN and activate the DMN [17]. A relaxed environment induces the brain to function in a "*mind wandering*" way, and problem solving is taken to an unconscious plane, in which automatic and spontaneous cognitive mechanisms take place. When this mode of operation is induced, the brain manipulates the information, adding, eliminating and transforming existing information, mixing it with random information, unrelated to the problem to be solved.

At a computational level this means that each agent in the group must have a different internal structure, and therefore they can manipulate the information in a different way. Alternatively, each agent can assume different roles, and therefore manipulate the information in a different way. In any case, agents must be able to remove or add information randomly, mixing it with the information that has been generated, allowing them to explore unexpected paths. In this way, novel and unexpected solutions can be obtained.

Therefore, if after a certain period of time, the MAS has not been able to solve a certain problem creatively, it can change its operating mode, assuming different roles and generating random information, mixing it with all the information that is it has accumulated throughout the process.

3. Short Duration

The long working sessions generate weariness in the participants and cause the participants to repeat over and over again established ideas. The reason is due to the fact that once a certain set of ideas has been generated, the participants tend to focus on their development rather than on the generation of new ones. Once certain search paths have been explored, certain neural connections are subtly reinforced, inducing the search paths associated with those connections to be explored again and again. For this reason it is preferable to do short sessions, and to continue new sessions after a certain time (for example after one or two days).

This way of functioning of the brain has been shaped throughout human evolution, since it guarantees our survival. The fact that humans continue to make the same conventional decisions that previously have been proven safe, has an evolutionary advantage over making continuous changes, and exploring new, unpredictable and less safe paths. Only when a strong conflict occurs, the brain stimulates the abandonment of certain patterns of activity and the adoption of new patterns, exploring new ideas.

Computer systems do not have these limitations so this aspect is not relevant for the design of a computer-based creative system.

VI. Guidelines for the Design of a Creative Computational System, Based on the Analysis of Brainstorming Variants

Although with an identical basic structure, many variations of Brainstorming methods have been proposed in order to increase their creative efficiency. Therefore, when analyzing each variant, certain complementary guidelines can be established, in order to optimize the structure of the MAS.

In general, the computational guidelines derived from the analysis of the different types of *Brainstorming* do not suggest the modification of the architecture of the MAS, and only suggest alternatives in its functioning structure, and the adoption of different roles by the different agents. Therefore, the "secretary agent" could make sure that the group of agents works in one way or another, according to the specifications shown in the different varieties of *Brainstorming*.

A. Stop-and-go Brainstorming

This method suggests that consciously delving into the problem moves in a certain direction, and the creative scene shifts to a new environment. In this new environment, ideas are generated again based on the initial approach and based on the information deduced by each participant of the group, which moves the scene to a new unpredictable creative environment, which contains the different perceptions of each component of the group.

This fact suggests that the computational system must alternate two types of thinking. One type of thinking is more spontaneous and random, and another type is more deductive and linear (problem solving).

B. Sequential Brainstorming

This method suggests that each participant goes deeper into a certain idea, since he accumulates all the information exposed by the others.

From a computational point of view the architecture of the system is the same and only the mode of operation varies. The different agents must be activated in an ordered sequential manner based on the information generated at each moment. In this way, an ordered list must be made with the order of activity of the different agents, in such a way that each agent will only be activated when the activity of the previous agent has finished. The work session will involve several cycles of activity of the different agents, and in each cycle the order of action of the different agents must be different. The "*secretary agent*" is in charge of deciding whether the MAS works in sequential mode, or in random mode.

C. Constructive-destructive Brainstorming

This method is very interesting and its effectiveness lies in stimulating creative ideas in an environment of reduced possibilities. The fact of generating destructive ideas in a first stage eliminates a priori certain conventional and easy search paths, in order to focus on the generation of ideas through other alternative and opposite paths.

From a computational point of view, this method can be carried out by initially changing, in the first operating cycles, the roles of the different agents, in order that they only generate destructive ideas, so that in a second place (in the following cycles of operation), change roles again, and engage in generating constructive ideas.

The "*secretary*" agent is in charge of deciding whether the MAS works in a *destructive mode*, or in a *constructive mode*.

D. Individual Brainstorming

This method does not suppose additional information when modeling a creativity computational system.

E. Anonymous Brainstorming

This method does not suppose additional information when modeling a creativity computational system. The only difference is that the way of acting of the agents is not conditioned by the proposals and behavior of other agents.

F. Brainstorming with Post-it (TM)

This method does not suppose additional information when modeling a creativity computational system. In a computer system, the activity of each agent does not have to be conditioned by other more prestigious agents (although in a conventional *Brainstorming* it can be programmed that some agents have a certain priority over others if that were the case). This activity can be done by the secretary at any time it is deemed necessary, since the secretary has information on the effectiveness of each agent and can provide them with different priorities.

Therefore, in a MAS architecture the different agents involved may have a similar weight, or on the other hand, in certain operating cycles the importance of the activity of certain agents can be weighted over others, in such a way that the information they generate is a priority (for the rest of the agents in subsequent operating cycles) over that generated by the other agents.

G. Brainstorming Phillips 66

This method does not suppose additional suggestions when modeling a creativity computational system.

H. Brainstorming Buzz

This method does not suppose additional suggestions when modeling a creativity computational system.

I. Didactic Brainstorming

This method is very interesting when modeling a creativity computational system since it seems to suggest that "*fuzzy*" searches should be created in parallel that are generally valid and none of these should be specified too soon. The information should be managed gradually and obtain general deductions without specifying. This aspect suggests that the "*creative group*" should work as a whole with various levels of abstraction, as it did with the "*problem solving group*". Each agent can act in parallel with various levels of abstraction, depending on the user's specifications, controlled, again, by the "*secretary agent*".

J. Brainstorming SIL (Successive Integration of Solutions)

This method is very suggestive since it proposes paths to follow in the process of developing ideas based on the forced grouping of previous ideas in an incremental way.

K. Brainstorming 635

This method is even more effective than the previous method, since it forces each participant to break the usual cognitive path several times, forcing them to explore lesser-known paths. The computational structure can be similar to that based on the previous method.

L. Brain Writing

This method does not suppose additional information when modeling a creativity computational system.

M. Collective Notebook

The fact of waiting one day between the generation of ideas sessions is based on the partial forgetting of the idea generated previously, so that the new idea arises from a collateral and non-evolutionary state. That is, the method aims to prevent it from deepening into a line of thought and instead tracked laterally. Therefore, it is an alternative to *Didactic Brainstorming*, and the same comments are valid to guide the development of a computational system.

N. Brainwriting Pool

The method is a new way of sharing previous ideas, although the computational structure may be similar to that of the previous methods.

O. Delphi Method

This method provides feedback to deduce new ideas. It is interesting and it is basic for a creativity computational system based on agents that interact with each other, feeding back a situation.

P. Nominal Group Method

This method is crucial, it is very effective since it generates many ideas and ensures that they are good without the need for a subsequent evaluation stage since the evaluation is continuous. As with the previous method, this method seems to suggest an agent-based system in which the group analyzes and advances the idea that each of them has generated to force again each of them to suggest a new idea based in those exposed as an extension of the one suggested previously.

VII. Guidelines for the Design of a Creative Computational System, Based on the Analysis of Graphic Brainstorming

Graphic Brainstorming [18] intends to facilitate the generation of graphic creations, such as architectural design, logo design, advertising image, fashion design, etc. This type of *Brainstorming* is complex, since the generation of abstract ideas must be complemented with the generation of graphic ideas, in the form of sketches, which represent the germ of the graphic composition that is intended to be achieved. Therefore, *Graphic Brainstorming* consists of at least three stages: idea generation, graphic generation and evaluation. And there are three methods to implement it:

- Shape Brainstorming. This method analyzes how parallels can be made between concepts and certain forms. For example: “designing a house with a rounded shape” would be something easy to specify. It is easy to find analogies between the circular concept and a circular shape to generate an initial action pattern. Instead, “designing a house with lots of light” would be much more difficult to establish initial formal analogies to generate initial sketches that can be refined in a later process. The key without a doubt is that the knowledge of things must be very extensive, and also ambiguous.
- Symbolic Brainstorming. This method is even more effective than the previous one, since the problem lies in finding a suitable symbol, but each symbol has a strong spatial and formal character, which can serve as a starting point for the design process. For example: “designing a very aggressive home”. The “aggressive” concept could be associated with a triangular or star shape, for example, or any existing symbol with many edges such as the triangular shape, the star shape, or similar. A computer system should make a great collection of previous associations between concepts and symbols, which means that a certain symbol should have a great quantity of attributes, and with ambiguous value.
- Metaphorical Brainstorming. This method is similar to the previous one, although more abstract, since the problem lies in finding a suitable metaphor, and each metaphor can again be associated with concrete or abstract forms, which can serve as a starting point for the design process. For example: “designing a house that stimulates spirituality”. The concept “spiritual” could be associated with common metaphors such as “ascension to heaven”, “communion with God”, etc. and these to forms usually associated with religions in a certain culture, such as a form of Latin cross, Greek cross, triangles, circles with radii, or a combination of them, for instance. A creativity computer system should make a great collection of previous associations between concepts and metaphors, which means that a certain symbol should have a multitude of attributes, and with ambiguous value.

The three *Graphic Brainstorming* methods suggest that both the “problem solving group” and the “creative group” should work at various levels of retraction, as already mentioned. This implies an important effort when it comes to representing knowledge, and when implementing the different processes for modifying, filtering and adding information.

VIII. Guidelines for the Design of a Creative Computational System, Based on the Analysis of WWW-Brainstorming

This method is a grouping of *Shape, Symbolic and Metaphorical Brainstorming*, but much more powerful, since the system incorporates a search engine for forms, symbols and metaphors on the web.

The analysis of the system suggests that the “creative group” of agents must have a specialized agent that looks for formal precedents, symbols, and metaphors in the web (therefore, it is called “www agent”). This search for associations in the web provides feedback to the system, and can be done at any time, although it is especially important at the beginning of the creative process.

This way of adding information to the system in order to stimulate new unexpected search paths is especially important, since the information that is added is not completely random, but has a certain type of connection with the problem to be solved.

IX. Guidelines for the Design of a Creative Computational System, Based on the Analysis of Lateral Thinking

Lateral thinking is a specific way of organizing thought processes, to find a solution through unorthodox strategies or algorithms, which would usually be ignored by logical thinking [19]. The mechanisms that promote *lateral thinking* have a direct influence on the conceptual architecture of the computational system to be designed. The strategies that promote lateral thinking are innumerable, and only the most frequent and effective are considered here. Other possible strategies do not have an impact on the design of the system, and only provide minor suggestions on the role of the different agents involved.

A. Random Words (Random Input)

A computer system does not have the cognitive restrictions of a human brain, so it is not difficult to break a certain logical structure of thought. The problem is deciding which logical structure to program. This technique suggests that a creative system should be based on two complementary and alternative problem-solving structures. First, a structure must be well defined, designed under intrinsic human parameters, and must be capable of solving a certain problem from a conventional point of view, in the way that most humans would, according to previously accepted conventional information and parameters. Another structure must be in charge of randomly breaking different processing stages of the previous system, introducing certain random information, more or less related to the problem to be solved.

Therefore, the analysis of this technique suggests that the creative computational system is based on a group of agents that can change roles when no truly creative ideas have been generated. That is, the group of agents with a “conventional role” would try to solve the problem. In the event that they cannot solve it, they would change their role, and adopt a “creative role” completely modifying their way of operating (in this case they can generate random proposals) and trying to reach a creative solution.

A more interesting alternative, which is emerging through this analysis, is that the MAS is made up of two groups of agents.

The "*problem solving group*" would try to solve the problem under conventional rational parameters that optimizes the obtaining of creative solutions in various ways. In the event that it cannot come up with any truly new solutions, it would cede control to the "creative group", whose agents have a completely different role, and in this case they can even come up with random ideas. In this way, the path for a solution seems wobbling as it is shown in Fig. 4.

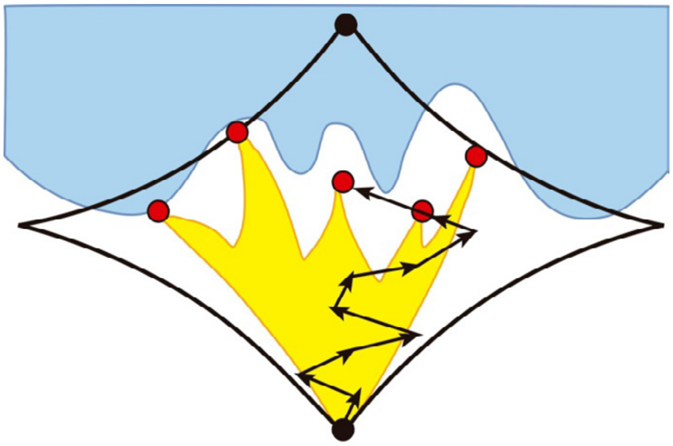


Fig. 4. Unpredictable path that allows one or more creative solutions to be achieved, using lateral thinking.

B. Delete Some Characteristic of the Problem

The analysis of this method suggests that the system must have a component that can manipulate the information within the "*creative group*" of agents. As it is explained in section X, in order to establish analogies at a high level of abstraction, the structure of the object-attribute-value information must have as many attributes as possible, and with the greatest possible variety of values. However, both the range of possible attributes and possible values must be continuously modified in order to find new exploration paths, and therefore obtain new ideas. The same can be said with respect to the characteristics of the problem to be solved. The problem to be solved can be altered, both in its initial definition and in the expanded definition that is completed as the resolution process progresses. This work could be done by an agent specialized in adding, filtering or modifying the information, the "*info-modifying agent*" at all times, under the control of the "*secretary agent*".

C. Modify or Exaggerate any Aspect Related to the Problem

The analysis of this method suggests, as it was the case with the previous method, that the "*creative group*" of agents should incorporate an agent ("*info-modifying agent*") in charge of modifying some attributes of the information related to the problem. This information can be altered, minimizing to exaggerating certain aspects of the information related to the problem. In this way, new paths can be explored, which if the information had not been modified would not have been explored, or simply nearby paths would have been explored.

D. Establish Analogies with Other Situations or Problems

As it has been said in the analysis of the establishment of analogies to solve problems, it is to base the establishment of analogies, at different levels of abstraction. Therefore, the information must initially be structured in several levels of abstraction, so that the "*creative group*" of agents can manipulate it at each different level in order to establish analogies with other previously solved problems.

E. Reverse the Problem

This method introduces distorting elements to avoid the chain of ideas that tends to be created in any conventional problem-solving system. Any algorithm is based, even if it is not desired, on certain chains of ideas, which often go unnoticed when programming. The same learning algorithms in reality what they do is a chain of ideas, it is just the opposite that is intended when implementing a creative system. In fact, the learning algorithms try to solve a certain problem based effectively, using the most appropriate paths in the decision tree, and therefore avoid exploring new paths.

The "*creative group*" should incorporate an agent ("*chunking agent*") that is exclusively dedicated to breaking the chain of ideas that tends to be repeated, and encourages the search for new paths, although initially they seem less promising. This agent must fragment the information usually united, and recompose it in different ways, although they may seem absurd. In this specific case we must reverse the problem. For example, when faced with the problem "designing a ship that rests on the sea", we could propose: "design a sea that rests on a ship". In fact, this idea was what drove the design of the hovercrafts.

F. Break the Problem into Different Components

This method is very useful, and in fact it is a basic strategy of the "*problem solving*" systems, called "divide and conquer". In this way to solve a problem it can be divided into parts, which can be solved separately, and to do so, each part can be divided again into parts that can be solved separately. At the end of the process, each sub-problem is so simple that it has an easy solution, and by concatenating all the solutions, a method is obtained to solve the general problem.

For example, to try to design a home, the kitchen, the bedroom, the living room, etc. can be designed separately, and in the same way, to design a kitchen the bench, the work area, the table, the furniture, etc. Once all the parts have been designed, the design of a house has been achieved.

In this sense, the "*problem solving group*" of agents should initially be able to identify parts of the problem, and once identified they can begin to be solved using a bottom-up structure. The problem with this strategy is that the link between all the parts must be established so that the whole is harmonious and well composed.

G. Take the Problem Out from Its Usual Context

By taking a problem out of its usual context, we are actually breaking pieces of information that are usually perceived as joined, and by doing so the pieces of information can be put back together in a different way.

Therefore, the "*chunking agent*" of the "*creative group*" must be able to establish random groupings of certain attributes of the problem to be solved with different values than the usual ones.

By randomly changing some of the initial conditions, the system would work in a strange and unpredictable way, would explore unsuspected paths and could arrive at some creative solution.

X. Guidelines for the Design of a Creative Computational System, Based on the Analysis of Parallel Thinking

Parallel thinking [20] allows contradictory opinions to coexist in parallel without having to be correct at every step, and in which there is no clash, no dispute, and no "true/false" initial judgment. In other words, this method basically indicates that a solution is good even though it has several adverse aspects.

Therefore, a computational approach would be based on the fact that some objectives delimited a priori could not be fulfilled and nevertheless take a certain solution as good. In this sense, several solutions could be found that, although they do not meet all the objectives, meet a minimum of essential basic objectives. However, these solutions are so novel that they can be taken for granted. Therefore, the "*secretary agent*" must be able to establish a hierarchy between the specified objectives, grouping them into indispensable, important and secondary. In the same way, the representation and manipulation of knowledge through *fuzzy logic* techniques must be endowed with a certain ambiguity.

XI. Conceptual Model to Implement a Creative Computational System, Based on the Analysis of the Main Methods that Stimulate Human Creativity

The suggestions made in the previous sections establish the principles for the design of the conceptual structure of a computational system capable of emulating the activity of the set of main methods that stimulate human creativity. The next subsections present the architecture for the information processing system and the information representation system.

A. Information Processing System

1. Definition of the Multi-agent Creative System

The collaboration of specialized agents organized as a MAS facilitates the implementation and integration of the methods and techniques suggested in the previous sections. This MAS is structured as two groups of agents: the "*problem solving group*" and the "*creative group*", as shown in Fig. 5.

The "*problem solving group*" has a basic structure containing at least two agents, the "*generator agent*" and the "*evaluator agent*". The problem solving group includes different algorithmic techniques of conventional AI, through which the tree of solutions is explored in search of adequate solutions. The only difference is that it can work at different levels of abstraction, and that is why systems based on the establishment of analogies take on a new creative character. Therefore, the generating agent can adopt several roles, and several levels of abstraction, while the evaluating agent simply limits itself to evaluating the possible solutions (both from the generating agent and the creative group) to accept one or more solutions.

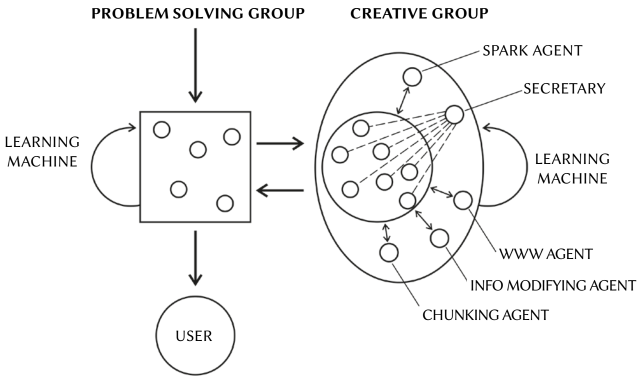


Fig. 5. Organization of a multi-agent creative system based on the analysis of the main methods to stimulate creativity.

The "*creative group*" is made up of a greater number of generic agents (between 5 and 7 agents), together with five specialized agents: the "*secretary*", the "*www agent*", the "*info-modifying agent*", the "*chunking agent*" and the "*spark agent*".

The "*secretary*" controls the operation of the "*creative group*", proposing the order of activation of the rest of the agents (sequential-random), changing their role (destructive-constructive), as well as their priority.

The "*www agent*" accesses the web on a regular basis, in order to find conceptual and formal analogies to the ideas generated by the group of agents in the group.

The "*info-modifying agent*" is in charge at all times of eliminating, adding or modifying the information related to all the possible objects involved in the resolution of the specific problem to be solved.

The "*chunking agent*" is exclusively dedicated to breaking the chain of ideas that tend to be repeated throughout the operation of the "*creative group*".

The "*spark agent*" is responsible for generating random ideas, or ideas contrary to those that may be generated at any time.

Both groups must have a "*learning system",* in order to learn from all the solutions generated, which ones have been finally chosen by the user, and they must know the differential reasons for which they have been chosen. The *learning system* has two components: the "abstract learning system", and the "details learning system". In order to avoid associations of a large number of pieces of knowledge, that is, in order to avoid the "chunking" of information that would force the system to repeat previously valid solutions, and therefore less creative.

The relationship between the different methods to stimulate creativity and the architecture of the system is as follows:

- *Establish analogies with known problems*. It is carried out by the "problem solving group"
- *Creativity matrix*. It is carried out by the "problem solving group"
- *Problem solving*. It is carried out by the "problem solving group"
- *Brainstorming*. It is carried out by the group of agents of the "creative group"
- *Variants of Brainstorming*. It is carried out by the group of agents of the "creative group"
- *Graphic Brainstorming*. It is carried out by the group of agents of the "creative group", in collaboration with the "www agent" and the "chunking agent"
- *www-Brainstorming*. It is carried out by the group of agents of the "creative group" in collaboration with the "www agent"
- *Lateral thinking*. It is done by the "chunking agent", the "info-modifying agent" and the "spark agent"
- *Parallel thinking*. It is done by the "chunking agent", the "info-modifying agent" and the "spark agent"

2. Functioning of the Multi-agent Creative System

In general, when the "*problem solving group*" is active, the "*creative group*" is not active, and vice versa, that is, they usually work in an antagonistic and complementary way.

Initially, the "*problem solving group*" is activated, and it will continue active until it finds a solution to a certain problem. The "*problem solving group*" can work at several levels of abstraction, therefore it can be activated in several sequential cycles, until the solution is properly specified.

If the group does not get a solution, or if the solutions it gets are not satisfactory, the "*problem solving group*" gives the control to the "*creative group*", which will be working until it achieves several solutions to the proposed problem. The list of possible solutions generated by the "*creative group*" is passed back to the "*problem solving group*" to be evaluated. Finally, the ideas that pass the evaluation process are finally presented to the user.

The "*creative group*" has a greater capacity to process information and has a greater number of agents, and can manipulate the information by adding information more or less related to the problem to be solved, and even random information. When adding, removing or modifying information, the rules that could be activated are different, so the troubleshooting strategy can vary considerably from one cycle to another, so the results are unpredictable.

B. Structure of Information Representation

The representation of knowledge is essential when designing a computational system that emulates the creative methods previously analyzed. In principle, the most appropriate representation is through the use of "object-attribute-value" structures for a specific object, idea or thought. This way of representing information has proven to be valid to represent even abstract and complex concepts, such as the existential rhetoric of architectural space [21].

First of all, the information on each object must be as extensive as possible, so it must have as many attributes as possible. In addition, it must have the largest possible number of relationship specifications with other objects, and it must have perfectly defined constraints. Finally, the range of possible values of attributes must be perfectly delimited, but with the greatest possible number of variations.

The greater the number of attributes and the greater the number of possible values for each attribute, the more creative the system can be. Let's take an example. A teacher does not have a ruler to draw a straight line on the board. And it will look at all candidate objects in the classroom that have a "weight" (very low, low, medium), a "length" (medium), and a "shape" (straight). Surely in the classroom there are candidates such as a drawer, a notebook, a tensioned cable, a computer keyboard and also a chair. The chair could have straight legs. In this way the teacher takes the chair and draws with it on the blackboard (see Fig. 6).

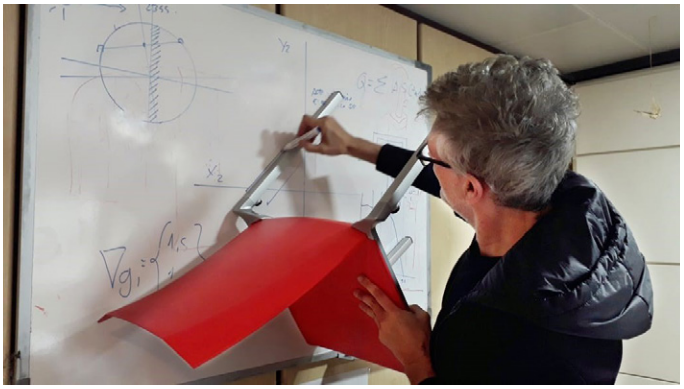
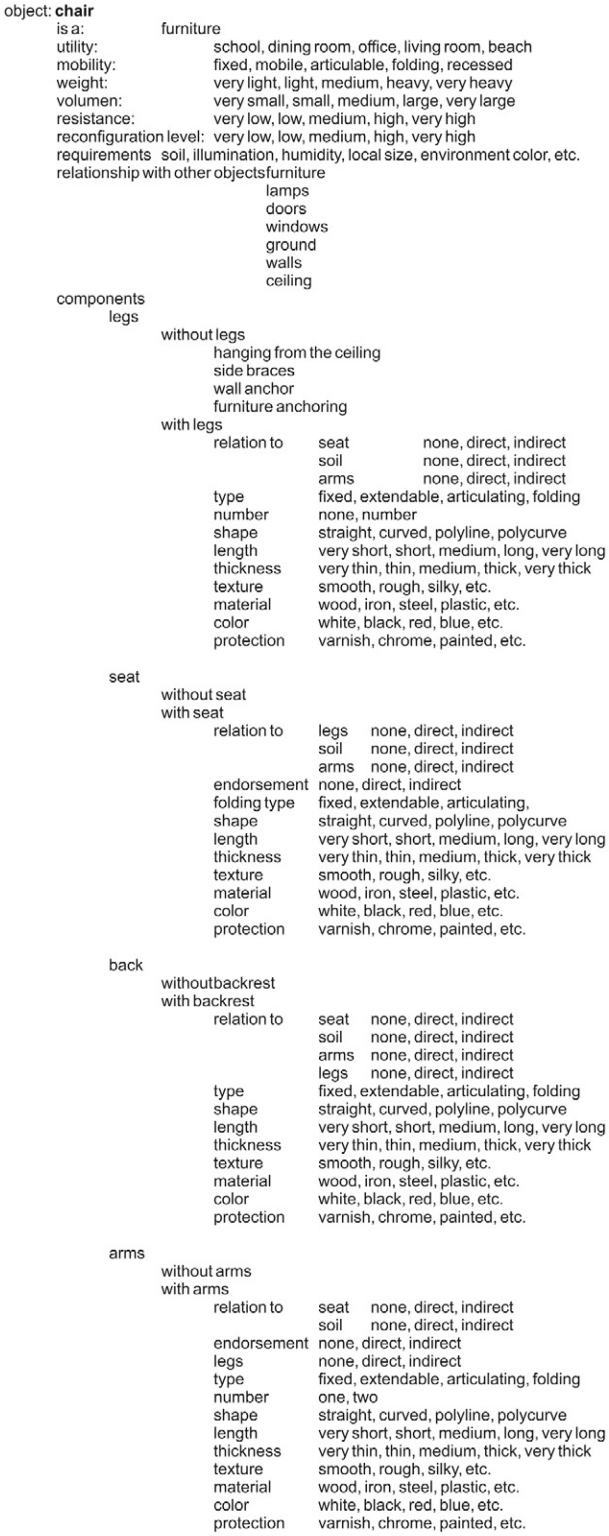


Fig. 6. Example of using lateral thinking for creative problem solving.

Students might think that their teacher is crazy (because he does unusual things) but also that he is very creative, or even a genius. The key is in how the knowledge of the objects in the classroom has been represented, and specifically the chair. For the teacher to use it to draw a straight line on the board, the chair must have the attribute "weight" (with a very low, low, and medium range of values) and the attribute "legs" with the attribute "shape" (with a range of values that includes the straight shape), and the attribute "length" (with a range of values that includes the average length).

Therefore, the initial representation of the information must be exhaustive, with the largest possible number of attributes, and with the largest possible range of values (see Fig. 7). In addition, due to the learning system, the information is gradually enriched with the activity of the system, being able to modify the structure of the information, the number of attributes and its range of values.



```
object: chair
          is a:               furniture
          utility:                      school, dining room, office, living room, beach
          mobility:                     fixed, mobile, articulable, folding, recessed
          weight:                       very light, light, medium, heavy, very heavy
          volumen:                      very small, small, medium, large, very large
          resistance:                   very low, low, medium, high, very high
          reconfiguration level:        very low, low, medium, high, very high
          requirements  soil, illumination, humidity, local size, environment color, etc.
          relationship with other objectsfurniture
                                            lamps
                                            doors
                                            windows
                                            ground
                                            walls
                                            ceiling
          components
                    legs
                              without legs
                                        hanging from the ceiling
                                        side braces
                                        wall anchor
                                        furniture anchoring
                              with legs
                                        relation to        seat          none, direct, indirect
                                                           soil          none, direct, indirect
                                                           arms          none, direct, indirect
                                        type               fixed, extendable, articulating, folding
                                        number             none, number
                                        shape              straight, curved, polyline, polycurve
                                        length             very short, short, medium, long, very long
                                        thickness          very thin, thin, medium, thick, very thick
                                        texture            smooth, rough, silky, etc.
                                        material           wood, iron, steel, plastic, etc.
                                        color              white, black, red, blue, etc.
                                        protection         varnish, chrome, painted, etc.

                    seat
                              without seat
                              with seat
                                        relation to        legs    none, direct, indirect
                                                           soil    none, direct, indirect
                                                           arms    none, direct, indirect
                                        endorsement        none, direct, indirect
                                        folding type       fixed, extendable, articulating,
                                        shape              straight, curved, polyline, polycurve
                                        length             very short, short, medium, long, very long
                                        thickness          very thin, thin, medium, thick, very thick
                                        texture            smooth, rough, silky, etc.
                                        material           wood, iron, steel, plastic, etc.
                                        color              white, black, red, blue, etc.
                                        protection         varnish, chrome, painted, etc.

                    back
                              withoutbackrest
                              with backrest
                                        relation to        seat    none, direct, indirect
                                                           soil    none, direct, indirect
                                                           arms    none, direct, indirect
                                                           legs    none, direct, indirect
                                        type               fixed, extendable, articulating, folding
                                        shape              straight, curved, polyline, polycurve
                                        length             very short, short, medium, long, very long
                                        thickness          very thin, thin, medium, thick, very thick
                                        texture            smooth, rough, silky, etc.
                                        material           wood, iron, steel, plastic, etc.
                                        color              white, black, red, blue, etc.
                                        protection         varnish, chrome, painted, etc.

                    arms
                              without arms
                              with arms
                                        relation to        seat    none, direct, indirect
                                                           soil    none, direct, indirect
                                        endorsement        none, direct, indirect
                                        legs               none, direct, indirect
                                        type               fixed, extendable, articulating, folding
                                        number             one, two
                                        shape              straight, curved, polyline, polycurve
                                        length             very short, short, medium, long, very long
                                        thickness          very thin, thin, medium, thick, very thick
                                        texture            smooth, rough, silky, etc.
                                        material           wood, iron, steel, plastic, etc.
                                        color              white, black, red, blue, etc.
                                        protection         varnish, chrome, painted, etc.
```

Fig. 7. Example of dynamic and exhaustive representation of information that allows the representation of information with various levels of abstraction, and the subsequent manipulation of the same by the "creative group" of agents. As a result, different representations (alternative and complementary) of the same object will coexist at each stage of the design process.

This general information is always available in the database, although each time the system tries to solve a certain problem it handles it differently. First, the system ("info modifying agent") classifies the attributes of the information based on their priority for a specific problem. Second, the system can filter the information in various ways. Third, the system can manipulate the information by adding information semantically related to the problem to be solved. Finally, the system can even add random information (see Fig. 8).

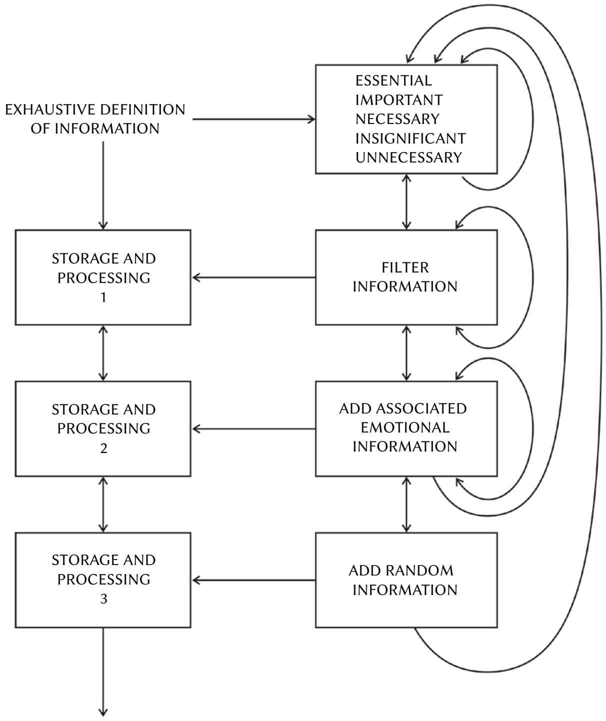


Fig. 8. Information processing system for the "creative group" of agents. The initial representation must be comprehensive and must be dynamically enriched through the learning system. However, the information can be modified continuously at each stage of the creative process, creating different, alternative and complementary representations of the same object.

The manipulated information can be temporarily stored during the resolution of a certain problem, but this does not alter the initial information available in the database at all times. Therefore the system has at all times several representations of the same object (called "virtual" representations).

In this way the agents can dynamically filter the information, making it "virtual", since at any moment the original structure of all the objects involved in the process of solving a certain problem can be recovered. In this way, the information available from the outside world is increasingly complex, with which the range of alternatives is increasing, and therefore the creative component of the system.

A key aspect is that the different attributes of objects must be organized, grouped and classified according to their importance. In this way, the dynamic and blurred manipulation of the information can be allowed, grouping the attributes in different ways or changing their priority at all times. This mechanism would allow the codification of the information storage and its dynamic and fuzzy retrieval, that is, the information, although apparently the same, would be reconstructed differently in each specific case. In this way, the fuzzy and dynamic mechanisms of human memory would be emulated.

XII. Discussion

The guidelines obtained in this paper can contribute to enhance existing works on the application of MAS for building creative systems. Some of the most relevant are reviewed below, with some indications on how our framework integrates and expands those ideas.

López-Ortega [22] considers deliberate creativity (the result of processing existing information), spontaneous creativity and human in the loop creativity (assisting a human operator in the creation act), and how agents can achieve them using planning, divergent exploration, selective attention, combination, and resolving. Creativity would result from the combination of these processes. Our work, however, goes further in the identification of processes and how they contribute to achieving creativity.

A model of organizational creativity [23] (i.e., the one organizations promote to survive producing new ideas, products, or services) proposes the collaboration of monitoring agents (checking the opinion of other agents), capturing agents (retrieving opinions from a database), and creative agents (retrieving ideas from a storage). Compared to our work, they address a limited number of creativity techniques, focusing in the identification of categories to evaluate the solutions.

Something in that line is the proposal by Macedo & Cardoso [24], who conceive the creation as a combination of three agents: the author agent (that produces the surprising products) and the jury-agents (that decide how surprising they are and can update the author agent emotional state towards its creation). They define the creation as a try-and-error process guided by utility functions, and the appraisal method. This author-agent and jury-agent scheme would be implicit in ideas of our framework, such as "eliminate criticism" (section V.A) to capture what the author agent does when the jury-agent does not show as much surprise as expected.

The SMART formal agent framework has been used to analyse creativity [25]. Creativity is seen as the result of autonomous interaction between parties and could be identified in different inherent activities to an agent framework, for instance in the perception of the agents, in the revision of models when taking decisions. Our approach considers the reverse, how creativity can be explicitly realised by agents by using human metaphors. These metaphors are important to understand what happens within the MAS and what the intended results are. After all, it is the observer who judges the existence of creativity.

Agents have been applied in the conceptualization of creativity in painting, with agents controlling fitness functions of ant-colony optimization software algorithms [26]. The work applies rather straightforward modifications of algorithms, letting the reader without an explicit guidance of why that particular approach makes sense. In this case, the creativity matrices (section III) would help explaining why some parameter modifications made sense.

Some works try to enhance individual creativity with assisting tools [27]. This is a P-creativity approach [3], where the tools present sources of inspiration and analogies to human operators, something in line with the ideas we presented in section II. That work did not propose a specific framework, unlike our work that bases upon a previous architecture with the goal of extending it with new methods. For instance, variations of brainstorming (section VI) could be applied to either generate analogies or new sources of inspiration (e.g. lateral thinking, Section IX).

Sosa and Gero [28] propose a socio-cognitive framework to analyse the interaction between designers and social groups to alter an artefact. The building block is a socio-cognitive agent, which has a first class representation of social norms. Agents playing the role of adopters (consumers of opinions) and promoters (opinion leaders) are embodying the idea of creativity in this construction. Their interaction propagates ideas on how to modify artefacts. Compared to their work, which performs simulations, we intend to produce real creative processes following the formulas expressed in the paper. These formulas applied to [28] would tell that further work could include vertical thinking techniques (section IV) to design ways to address modifications in the artefacts. Also, additional variations could be generated using creativity matrices, since they permit to draw analogies and derive variations (section III).

MASTER (Multi-Agent System for Text Emotion Representation) is another experiment where each agent has a digital emotional state and can influence others by reciting poems [29]. The work focuses on the poetry generation domain and incorporates techniques to generate poems, but also to interpret them. The reaction of the listener gives clues to the speaker to alter the poem and get a higher effect. Our approach is more generic, although our first experimentation focuses on the drawing domain. Unlike [29], we have not defined specific semantics, but this allows for generalizing the results to other problems. The scheme applied in [29] could be enriched with more tactics, such as generating lots of poems to have more feedback (section V.C) or the multiplying effect if a poet agent reuses successful poems for a number of agents (section V.D).

Mendez et al [30] use agents to create stories. There are director agents, to direct the story plot, that create new character and object agents, set the motivation for characters, and take care that characters do not perform undesired actions. Character agents use affinity models to regulate interactions between them. Our framework captures this process as a brainstorming, maybe a stop-and-go brainstorming (section VI.A) because the director agents puts barriers to how the story progresses and creates shifts in the plot.

XIII. Conclusions

In this work, the most important techniques that stimulate human creativity have been identified and analysed. Based on this analysis, the corresponding and appropriate parallels have been suggested for the conceptual design of an agent-based creative system. This system considers two groups of agents, one that uses some conventional problem solving techniques, and a second one that generates new paths by using creative methods, which are inspired by those used by humans. These new paths can be then executed and evaluated by the problem solving group in order to determine whether they drive to effective and potentially creative solutions that can be shown to the user.

A first prototype of a MAS using these ideas was reported in [1], by using the INGENIAS agent methodology [31], although it does not implement all the features defined in this paper. That MAS is being developed for specific case studies (e.g. the design of a chair), and assisting the user in their creative process. It starts with the identification of the knowledge representation model of the domain, with attributes and relationships as shown in the example above (see Fig. 7). This can be supported by the use of ontologies. As well, the agents in the knowledge processing system requires, besides the internal behaviour of each type of agent, the organization and information of them.

An issue that has not been developed yet is the way of determining the degree of creativity of the solutions generated by the system. This is questionable because it is already difficult (if not impossible) to determine it for human creations (this has been already discussed in several works, as it is mentioned in the introduction). It would be possible to define some metrics for it [32], and these can be checked by a new group of agents, in a similar way to some approaches that have been discussed in section XII. However, it is possible to determine whether the generated solutions satisfy requirements that define the problem, and this is the purpose of the problem solving group in our framework. At the end, the human being watching the results will determine which ones like more or consider more creative.

Acknowledgements

This work has been partially supported by the project "Collaborative Design for the Promotion of the Well-Being in Inclusive Smart Cities (DColbici3)" (grant TIN2017-88327-R) funded by the Spanish Ministry for Economy, Industry, and Competitiveness.

References

[1] L. De Garrido, J. Gomez Sanz, J. Pavón, "Agent-based Modeling of Collaborative Creative Processes with INGENIAS," *AI Communications*, vol. 32, no. 3, 2019, pp. 223-233.

[2] M. Boden, "Creativity and artificial intelligence", in *Artificial Intelligence*, vol. 103, no. 1-2, 1998, pp. 347-356.

[3] M. Boden, "Creativity," in *Artificial Intelligence,* Academic Press, 1996, pp. 267-291.

[4] K. Jennings, "Developing creativity: Artificial barriers in artificial intelligence," in *Minds and Machines*, vol. 20, no. 4, 2010, pp. 489-501.

[5] R. A. Brooks, "Intelligence without reason". Massachusetts Institute of Technology. *Artificial Intelligence Memo No. 1293*, 2010.

[6] M. Rhodes, "An analysis of creativity," in *Phi Delta Kappan*, vol. 42, no. 7, 1961, 305–310.

[7] A. Jordanous, "Four PPPPerspectives on computational creativity in theory and in practice," in *Connection Science*, vol. 28, no. 2, 2016, pp. 194-216.

[8] C. Lamb, D.G. Brown, and C. Clarke, "Evaluating computational creativity: An interdisciplinary tutorial," in *ACM Computing Surveys* vol. 51, no. 2, 2018, pp. 1-34.

[9] G. A. Wiggins, "A preliminary framework for description, analysis and comparison of creative systems," in *Knowledge-Based Systems*, vol. 19, no. 7, 2006, pp. 449-458.

[10] M. Martin, and K.I. Voigt, "What Do We Really Know about Creativity Techniques? A Review of the Empirical Literature," *The Role of Creativity in the Management of Innovation*, 2017, pp. 181-203.

[11] S. A. Leybourne, "The Creativity Matrix: Balancing Architectural and Process Creativity in Project-based Management," in *IRNOP Conference*, Oslo, Norway, 2013, pp. 17-19.

[12] V. Tang, J. Luo, "Idea matrix and creativity operators," in *DS 75-7: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol.7: Human Behaviour in Design*, Seoul, Korea, 2013, pp. 19-22.

[13] M. A. Runco, *Problem Finding, Problem Solving, and Creativity*, California, USA, Greenwood Publishing Group, 1994.

[14] I. Wopereis, E. Derix "Seeking Creativity: A Case Study on Information Problem Solving in Professional Music," in *Kurbanoğlu S. et al. (eds) Information Literacy: Key to an Inclusive Society, ECIL 2016, Communications in Computer and Information Science*, vol. 676, Springer, 2016.

[15] J. G. Rawlinson, *Creative thinking and brainstorming*, Routledge, 2017.

[16] A. F. Osborn, "Creative Thinking," in *American Association of Industrial Nurses Journal*, vol. 6, no. 9, 1954, pp. 23-25.

[17] M. E. Raichle, "The brain's default mode network," in *Annual Review of Neuroscience*, vol. 38, 2015, pp. 433-447.

[18] L. De Garrido, *Applications of Artificial Intelligence in the composition of architectural objects.* PhD Thesis, Escuela Técnica Superior de Arquitectura, Universidad Politécnica de Valencia, Spain, 1989.

[19] E. De Bono, *The Use of Lateral Thinking*, London, Jonathan Cape Ltd. 1967.

[20] E. De Bono, *Six thinking hats*, London, UK, Penguin, 2017.

[21] F. Garijo, and L. De Garrido, "A knowledge based system for house design," in *Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics*, 1988, pp. 806-809.

[22] O. López-Ortega, "Computer-assisted creativity: Emulation of cognitive processes on a multi-agent system," in *Expert Systems with Applications*, vol. 40, no.9, 2013, pp. 3459-3470.

[23] C. Olszak, and T. Bartus, "Multi-agent approach in designing of organizational creativity support," in *The European Conference on Information Systems Management*, Academic Conferences International Limited, 2015, pp. 93.

[24] L. Macedo, and A. Cardoso, "Using Surprise to Create Products that get the Attention of other Agents," in AAAI Fall Symposium, 2011, pp. 79-84.

[25] M. d'Inverno, and M. Luck, "Creativity through autonomy and interaction," in *Cognitive Computation*, 4(3), 2012, pp. 332-346.

[26] G. R. Greenfield, "Computational aesthetics as a tool for creativity," in *Proceedings of the 5th conference on Creativity & cognition*, 2015, pp. 232-235.

[27] U. Lösch, J. Dugdale, Y. Demazeau, "Requirements for supporting individual human creativity in the design domain," in *International Conference on Entertainment Computing*, Springer, 2009, pp. 210-215.

[28] R. Sosa, J. S. Gero, "A computational framework for the study of creativity and innovation in design: Effects of social ties," in *Design Computing and Cognition'04*, Springer, 2004 pp. 499-517.

[29] A. Kirke, E. Miranda, "Emotional and multi-agent systems in computer-aided writing and poetry," in *Proceedings of the Artificial Intelligence and Poetry Symposium*, 2013, pp. 17-22.

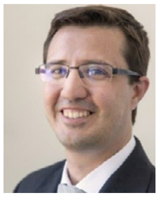
[30] G. Méndez, P. Gervás, and C. León, "A model of character affinity for agent-based story generation," in *9th International Conference on Knowledge, Information and Creativity Support Systems*, Limassol, Cyprus, vol. 11, 2014.

[31] J. Pavón, J. J. Gómez-Sanz, R. Fuentes, "The INGENIAS methodology and tools," in *Agent-oriented methods*, IGI Global, Idea Group Publishing, 2005, pp. 236-276.

[32] J.J. Shah, S.M. Smith, N. Vargas-Hernandez, "Metrics for measuring ideation effectiveness," in Design Studies, vol. 24, no. 2, 2003, pp. 111-134.

Luis De Garrido

Architecture Department. Universitat de València. Spain. National Association of Sustainable Creative Architecture Research. Spain. PhD degree in Architecture (Universidad Politécnica de Valencia, 1989) and Doctor Honoris Causa (Universidad San Martin de Porres, 2019). Currently he develops a professional activity as an architect, professor and independent researcher.



Jorge J. Gómez Sanz

Institute of Knowledge Technology. Universidad Complutense de Madrid. Spain. Associate Professor in the Universidad Complutense de Madrid. He holds a degree in software engineering and a PhD (2002). His PhD introduced a new methodology for the development of Multi-Agent systems: INGENIAS. He has focused his research in agent oriented software engineering.

Juan Pavón

Institute of Knowledge Technology. Universidad Complutense de Madrid. Spain. PhD degree in Computer Science (Universidad Politécnica Madrid, 1988). Currently he is Full Professor at Universidad Complutense Madrid, where he leads the GRASIA research group and the Institute of Knowledge Technology. His main areas of interest focus on the application of Artificial Intelligence in multidisciplinary projects with social value, such as assistive technologies, health monitoring, ambient assisted living, smart cities, education, computational creativity, and tools to support Responsible Research and Innovation (RRI).

Towards a Solution to Create, Test and Publish Mixed Reality Experiences for Occupational Safety and Health Learning: Training-MR

Miguel A. Lopez[1], Sara Terron[1], J. M. Lombardo[1], Rubén Gonzalez-Crespo[2] *

[1] Fundación I+D del software libre (FIDESOL), Granada (Spain)
[2] Universidad Internacional de La Rioja, Logroño (Spain)

Received 29 April 2021 | Accepted 1 July 2021 | Published 23 July 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

Artificial intelligence, Internet of Things, Human Augmentation, virtual reality, or mixed reality have been rapidly implemented in Industry 4.0, as they improve the productivity of workers. This productivity improvement can come largely from modernizing tools, improving training, and implementing safer working methods. Human Augmentation is helping to place workers in unique environments through virtual reality or mixed reality, by applying them to training actions in a totally innovative way. Science still has to overcome several technological challenges to achieve widespread application of these tools. One of them is the democratisation of these experiences, for which is essential to make them more accessible, reducing the cost of creation that is the main barrier to entry. The cost of these mixed reality experiences lies in the effort required to design and build these mixed reality training experiences. Nevertheless, the tool presented in this paper is a solution to these current limitations. A solution for designing, building and publishing experiences is presented in this paper. With the solution, content creators will be able to create their own training experiences in a semi-assisted way and eventually publish them in the Cloud. Students will be able to access this training offered as a service, using Microsoft HoloLens2. In this paper, the reader will find technical details of the Training-MR, its architecture, mode of operation and communication.

Keywords

Human Augmentation, Mixed Reality, Virtual Reality, Education, Elearning.

DOI: 10.9781/ijimai.2021.07.003

I. Introduction

New technologies are transforming the society in which we live at a breakneck pace. The application of new technologies takes place in various areas in order to improve productive processes, facilitate personal relationships, or help to better understand society. This global transformation reaches its full potential in transforming the more traditional industry into the new 4.0 industry which is already a reality [1] [2]. Sensors networks, embedded systems, or wearable devices networks are interconnected to form large IoT networks (Internet of Things). These IoT networks have enhanced the interoperability of companies [3]. In addition, new artificial intelligence technologies have been incorporated to a large extent in the industry, empowering the use of data to optimize, automate and improve various types of processes [4]. This ecosystem is in continuous technological revolution, which encourages other scientific fields to be growing strongly. Technologies related to “Human Augmentation” (HA) seek to offer technological solutions to improve people’s productivity by using different tools and algorithms. A subset of these tools are those encompassed by the "continuum reality" [5]. This concept proposes other realities that are accessible through the use of new technologies. In this way, users can experience situations and perform actions different from the real ones, depending on the position in which we are in the "continuum reality". Within this area of research, virtual reality is experiencing an exponential growth in recent years [6]. The possibility of placing the user in a controlled, completely real and highly interacting environment has encouraged many researchers to explore the applicability in engineering [7], medicine [8], or education [9], among others. In this case, users interact with virtual elements. Virtual reality is not the only alternative explored by researchers. The rise of smartphones with high computing power and a camera with appropriate technical features, provides a perfect platform for the execution of many augmented reality solutions [10]. Digital elements are represented in the user’s visible spectrum through an external element. These elements can interact with each other or as a result of user actions. The development of the latest hardware platforms by the large manufacturers of the technology industry has been made possible by mixed reality to assist in software development, which have been applied in industry, architecture, engineering or construction [11] [12]. For all the so-called extended realities, researchers have studied how these technologies could be applied to training and education, in order to enhance students’ performance and skills. With regard to training, particularly for industry, the focus on occupational risks

* Corresponding author.

E-mail addresses: malopez@fidesol.org (M. A. López), sterron@fidesol.org (S. Terron), jmlombardo@fidesol.org (J. M. Lombardo), ruben.gonzalez@unir.net (R. Gónzalez-Crespo).

prevention should be highlighted. Occupational Safety and Health (OSH) is a major challenge for society and science. In 2016, about 3 million workers in the industry sector reported an occupational injury or illness, which is equivalent to 2.9% of full-time workers registered in the United States [13]. The science still has a long way to go to clearly and correctly identify the factors that can cause occupational accidents or professional illnesses. Within the scope of interest of this paper, we find research related to the use of extended realities applied in the training of workers, such as virtual reality or mixed reality. Specifically, we propose our solution, whose objective is the democratization of training with extended realities through technologies applied to the prevention of occupational risks. This solution offers a cloud service for terminals of different technologies with which students may experience situations of risk, but without compromising their physical integrity or implying any cost in materials, and without causing possible damage to the company's facilities or resources. This type of student-centered training will help to improve their rapid response to emergency situations, as well as to know the protocols to be followed in order to carry out the work in appropriate conditions of safety and health. The training activity can be performed in a delocalized way with virtual reality or on site with mixed reality. Furthermore, to break the barrier of the cost of applying this technology [14] [15] [16], our development offers a set of tools for the creation of training experiences. A trainer can use these tools to build their own mixed reality experiences in a completely customized way depending on the workplace.

This article is structured as follows: the background when the authors introduce the library review. Then, in motivation and methodology we present the result of the OSH analyst and the Training-MR objective. We continue with the technical description where we resume the main issues, details, characteristics about Training-MR. At the end, discussion and conclusion are presented where we analyse the advantages of the Training-ME and present its limitations and future lines.

II. Background

Human augmentation comprises a field of science whose objective is to improve human capacities through the use of tools, which can have a different degree of integration with people's actions and perceptions of their environment. A common example of these tools is the devices used by people with reduced hearing capacity. Human augmentation can be differentiated from other similar fields of research such as Human Enhancement (HE), in which the improvement of the human body itself is pursued. That is, HE seeks to improve the human body through the use of various technologies [17], while HA focuses on the application of technologies to improve human capacity and productivity, without the need to modify the body itself. An example of this differentiation is found in the current use of mixed reality glasses that provide real-world digital information, in front of a hypothetical artificial eye that sends digital and real information to the user's brain. In this case, the use of mixed reality glasses corresponds to HA and the artificial eye with HE. Providing a clear definition for HA is not easy, so several definitions of this concept can be found in the scientific literature [18]. Li introduces human augmentation technology referred to "methods with which human beings can obtain abilities exceeding the normal level or can compensate for abilities impairments" [19]. Another main definition is that provided by Rasiano [20]. In their study it is presented as "an interdisciplinary field that addresses methods, technologies and their applications for enhancing sensing, action and/or cognitive abilities of a human. This is achieved through sensing and actuation technologies, fusion and fission of information, and Artificial Intelligence methods". The study of HA is often divided to improve its understanding and study. Li proposes a classification according to the scientific field and the impact of the adopted technology on the user. The four categories identified are as follows [19]:

- Medication augmented: for research focusing on the use of medication.
- Genetic augmented: for research using genetic modification techniques.
- Mechanical augmented: for research that proposes the use of hardware or electronics.
- Surgical augmented: for research focused on surgical operations of patients.

Other categories of HA have been proposed, such as Rasiano's research explaining three categories according to augmented skill [20]:

- Sense augmented: improvement of the user's ability to perceive the world.
- Action augmented: improvement of the user's performance capabilities.
- Cognition augmented: improvement of the cognitive abilities of the user.

Based on the numerous considerations that can be found in the literature and the studies carried out in this field, we define Human Augmentation as "the augmentation of the user from devices with which they are equipped or dressed, to improve the results of tasks by transforming the way they are performed". Thus, a categorization of the human augmentation is proposed with an approach that is not based on the augmented user capacity or the way in which this is achieved. Today, and with a foreseeable increase in the future, this augmentation will be achieved by several means (Li's proposal [19]), and will affect several capacities (Raisamo's proposal [20]). Accordingly, it is already common to find a single device that influences both senses and actions (the Microsoft HoloLens, for example). Taking this into account, we offer our own categorization based on the augmentation achieved by the user from the technology implemented for this purpose:

- Augmented successfully: the application of one or more HA technologies enable a user to perform tasks that would otherwise be unfeasible, or even achieve a more efficient performance. An example is the combination of mixed reality glasses with hand detection devices and a remote robot to perform underwater operations.
- Augmented Multitasking: the HA allows a user to perform parallel operations, which would otherwise have to be done sequentially. This would be the case, for example, of executing a complex task in an industrial production chain with two collaborative robots, one operated by voice commands and the other by hand gestures.
- Augmented perception: set of HA devices that provide information to the user about the environment around them. This category groups all technologies focused on new design, creation or research modes, in which information and data are the main object of actions. For example, a scientific experiment conducted with mixed reality simulations.

Focusing on the application of these technologies in the field of occupational safety and health, some studies can be found.In 2012, the EUSafe project was created in Europe to promote the study on the prevention of occupational risks, supporting research in this área [21]. In this regard, education and training of all the roles involved in the occupational risk prevention chain plays a key role, including from auditors and inspectors in preventive matters to the workers themselves [21] [22]. In response to the needs of society and international research, a number of new technologies have

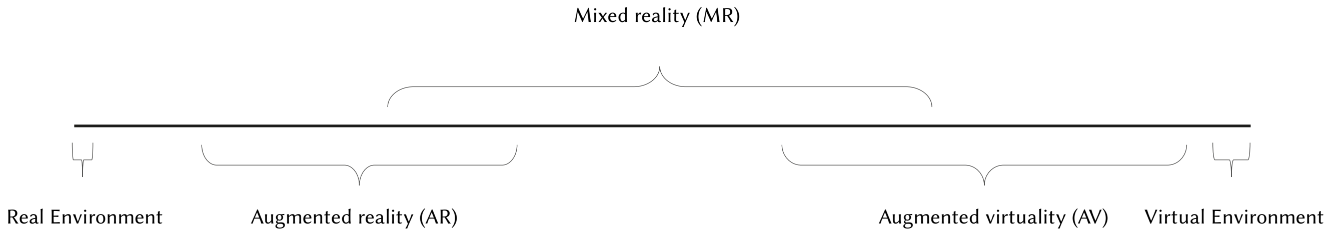


Fig. 1. Virtual continuum schema from Milgran y Kishino [5].

been successfully applied. Using artificial intelligence, we find the study of Simeone et al. in which they address the construction of a cloud platform for monitoring workers, in order to avoid accidents or injuries resulting from their usual work [23]. The application of Internet of Things (IoT) technologies and principles for the protection of workers also corresponds to another well-explored scientific area. Sensorizing workers' behavior has benefited workplace safety, as research by Suganya et al. shows, which exposes the construction of a miner monitoring system that includes various personal sensors and equipment to ensure their safety [24]. Another IoT study for the protection of workers in coal mines is conducted by Kumar et al. [25]. One of the main applications of IoT technologies in this area focuses on the helmet of workers. This individual protection element has unique characteristics (type-approval, obligation to use, position and guidance vis-à-vis the user). Thus, these protection elements become an integration hub for IoT sensors and devices deployed on the workers themselves [26] [27].

Considering the technologies that the HA encompasses, this paper focuses on the extent of reality, so we aim to improve the user's ability to perceive and interact with the surrounding environment. There are several types of applications of these realities, which depend on the digital tools selected. In this regard, Milgran and Kishino proposed the so-called "virtual continuum" [5]. This "virtual continuum" corresponds to the linear representation presented in Fig. 1, in which the technology is located according to its proximity to the real world (free of digital elements), or virtual (where every user-perceptible element is digital). There is a very wide range of possibilities between the two ends of the line. Augmented reality is the exposure of user-perceptible digital elements within their environment. An example of everyday use can be a vehicle browser application that expands the user's perception of his environment, providing the user with the appropriate direction. On the other hand, the increase in the load of digital elements while reducing the user's perception of their real environment, lead to the approach towards extended virtual reality. Mixed reality, meanwhile, occupies a distinct position as it displays digital 3D elements without removing user perception from the real environment. For the purpose of this study, we consider two positions within the "virtual continuum" to be of interest. The first of these is the virtual reality, positioned on the far right of the image, corresponding to the virtual environment, in which only digital elements are perceived by the user. The user feels immersed in a virtual environment. The second is the mixed reality, in which users visualize and interact with digital elements while still perceiving the real world. In mixed reality, digital objects can interact with each other or with other objects in the real world.

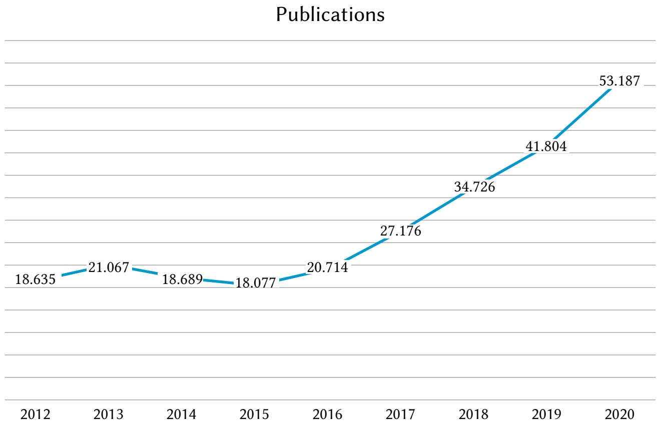


Fig. 2. Chart of papers with virtual reality key word. Data from https://app.dimensions.ai/.

A. Virtual Reality

Virtual reality (VR) was introduced in the 1960s, experiencing various modifications over the years due to the advancement of science and technology. Gigante in 1993 identified virtual reality as "the illusion of participation in a synthetic environment rather than external observation of such an environment. VR relies on three-dimensional (3D), stereoscopic, head-tracked displays, hand/body tracking and binaural sound. VR is an immersive, multisensory experience" [28]. On the other hand, in 2019 Kardong-Edgren states the need to agree on a term that today meets the variety of applications and research carried out around this concept [29]. The research in HA and VR is currently at a time of great relevance for their research. Technological progress has led to the lower price of technology, which has turned VR glasses into a common consumer product for the general population. This has influenced the conduct of applicability studies in different fields of research and the consequent increase in the number of scientific publications. As Fig. 2 shows, the number of scientific papers containing the keyword "virtual reality" continues to grow over the years. The increase in published scientific work reflects the opportunities offered by these technologies. Isolating a user in a simulated, virtual, and 3D environment from the real world is a useful tool. Also, it is worth noting the degrees of freedom available to a user in virtual reality. The most commonly used devices such as Oculus Quest or HTC VIVE are offered as elements of interaction with the virtual environment. The set of "head mounted displays" (HMD) and hand tracking devices in VR offer the user freedom related to: (1) head tracking, i.e. the user's view in the real environment is driven by the movement of their head; (2) wrist rotation, so that virtual hands rotate with the movement of the user's wrist; (3) head and hand orientation are interpolated to estimate a natural position of the arms, although in most virtual environments the user's arms are removed from simulation; (4) the movement of three of the five fingers of each hand are detected by virtual reality hand tracking devices and sent to the digital fingers, interpolating the remaining two fingers to provide a feeling close to reality. It is common for current HMD devices to have a way to recognize the depth of field in a particular configured environment, which in Oculus Quest is called The Guardian. This provides the user with a controlled environment where their position is detected at all times, giving freedom of movement, also taking into account their height to know if they are crouched or standing. Thus, with modern systems, a user in a virtual world can simulate a controlled movement in a given space and even interact through the height of his gaze, detected by the movement and position of the head. As far as the interaction with the hands is concerned, the sensation experienced is quite similar to the reality, since the most common movements that a person makes, such as picking, pressing, releasing or pinching with his thumb and forefinger, are perfectly detected. These virtual immersion features have popularized its use. The application of virtual reality in medical research has been very relevant, given the trajectory of this technology for simulation. Quigley et al. conducted a study on the use of VR for the training of patients aimed at weight loss [30]. Lombardo and López, both authors of this paper, also explored the application of VR in support of the rehabilitation of Parkinson's patients [8]. Similarly, many studies highlight the pedagogical virtues of the use of these tools [31] [32] [33] [34]. Different levels of application of virtual reality in the education and training of people can be found. A first level is identified with the implementation in basic education [35]. In more advanced studies such as those at the university, we find the research of Porter et al., in which they concluded that the use of VR improved the performance of the students, compared to those who were trained using only books or videos [36]. Du et al. performed a study with two different types of training experiences in VR, in which one group of students participated in the virtual experience individually, while in the other group more than one student was connected in the same experience and could interact with each other [9]. The results showed that students who participated in both types of training experiences with VR obtained better qualifications than those who used traditional methods. It is not difficult to find in the scientific literature studies that apply virtual reality in training oriented to the industrial sector, given the virtues of its adoption. VR has been used in different sectors to train professionals in the execution of tasks where their integrity and health may be at risk. As an example of this kind of training, we can cite the one associated with the firefighting. Among these studies were those carried out by Rahmalan et al. who used VR to instruct in estimating a fire [37], the one by Pitana et al. focused on training fire inspectors [38], or the study conducted by Wan for training inspectors of industrial oil deposits [15]. In line with the above studies, Li explored the use of VR for coal miners training [39]. Likewise, the authors of this paper have already introduced a new system to bring virtual reality closer to teachers [40], and have explored the use of a semi-assisted virtual experience creation system for training in the prevention of occupational risks [41]. After a thorough analysis of the state of the art, the following conclusions should be highlighted:

- Virtual Reality is a tool with a wide range of industrial applications.
- Virtual Reality can be successfully used for professional training so that they can act in situations of risk without affecting their health or physical integrity.
- The creation of all virtual experiences starts from scratch without using standard tools or framework, which implies cost overruns in the design phase.
- Existing solutions do not take into account the vulnerability of data that users expose in the system. Users in a virtual environment are providing information about themselves and how they interact, so these data must be properly protected.

B. Mixed Reality

As we saw in Fig. 1, another position of the "virtual continuum" is occupied by mixed reality (MR), whose characteristics are of interest to the object of this paper. The user in this region can perceive real and virtual objects, together but distinguishable from each other. Virtual objects must also interact with real ones. Also, the user has the ability to interact with these virtual objects in a natural way. For example, in a mixed reality scene that simulates the passage of objects through a real production chain, these objects must replicate real-world behavior and interact with both the user and the actuators in the assembly chain. Defining mixed reality is not an easy task, because of its constant evolution [42]. Milgran and Kishino defined it as "a mix of real and virtual objects within a single display." [5]. In 2019, Speicher et al. conducted a bibliographic review in order to provide a more specific definition of mixed reality, but their conclusion was that it "depends" [42]. Even if consensus has not been reached on the definition of MR, it is important to note that there is a difference from augmented reality (AR). Some authors describe the MR as an integration of VR and AR. In their case, Tepper et al. noted: "mixed reality merges many of the benefits of virtual reality and augmented reality" [43]. In other words, they offer the capabilities of a virtual world, where everything that happens is controlled by software, along with the user's perception of the real world. Analyzing the devices currently available on the market, the most commonly used is the Microsoft HoloLens device. Hololens is a Head Mounted Display (HMD) so it is placed on the head without the need to hold other devices on the hands. Version 2 is currently on the market. These devices offer a range of possibilities to define the experiences available to users. The main features of HoloLens are [44]: (1) head and eye tracking system; (2) microphone for voice commands; (3) accelerometers for the user's motion control, based on the acceleration of the head; (4) hand and finger tracking by computer vision and depth sensors. Also noteworthy are the actions that can be achieved by the use of hands. Microsoft Hololens has a very efficient gesture recognition system (e.g. hand closure, thumb grip and index finger, or select using index finger) (Fig. 3).

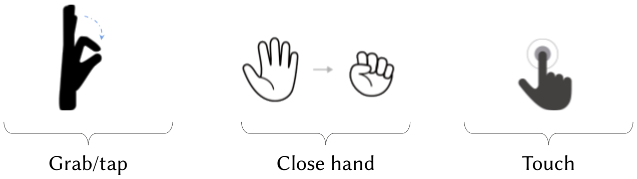


Fig. 3. HoloLens 2 gestures examples [45].

However, the HoloLens device also has limitations, as do any technology. For gestures, content creators should build virtual experiences according to the Hololens detection system, since hands can be hidden when using cameras and therefore their actions will not be detected. Research on virtual reality has accelerated the application of mixed reality in different cases of use. However, it is important not to consider that MR is simply an evolution or improvement of AR. This confusion may stem from the fact that the MR is a post-AR technology, but MR is really a technology that explores a different position from the "virtual continuum". Mixed reality makes digital information ubiquity possible, which has led to a significant increase in its application in the industry [46]. Previous research in augmented reality has facilitated the rapid reception of the MR. This is because some limitations of AR have been corrected by the new capacities provided by the MR [47]. The main applications of this technology in the industry support work at different points in the production chain. This support has mainly involved the incorporation of digital information into the execution of tasks [48]. Some of the most interesting applications are those related to the design of vehicles using aggregated information [49]. Another example is the adoption of the MR to offer operators a new user interface while working alongside robots or remotely with complex devices [50] [51]. It can be noted that significant efforts have been made in the scientific community to integrate the MR into the aviation industry, with particular emphasis on its airplane maintenance lines [52] [53] [54] [55] [56] [57]. The voice command interaction provided by the Hololens [58], together with the ability to provide information and designs at the operator's workplace, makes the MR a key tool for productivity improvement in the coming years [49] [59] [60], as well as for design, maintenance, security and quality control in the industry [48]. Likewise, the ability to interact and expose digital information in a real environment has made the MR a tool of relevance in the field of education and training in the industry [60]. The MR has been used in training to learn how to react in situations of risk [61] [62], or to provide a training environment for the performance of real tasks, greatly enriching the traditional training experience [63]. This improvement is seen in the new ways of training health care workers, taking advantage of the mixed reality both in tasks of special difficulty and in other everyday tasks such as stitching a patient [64]. These new VR and MR tools offer an opportunity to evolve educational and training methodologies [65]. Therefore, the MR is a tool that can be used in many areas such as industry, education, medicine, etc. The adoption of these new tools requires a proper process, since interaction with these elements may not be simple and therefore be rejected by users. For this reason, a process of user training must precede any action to implement a MR tool [48]. In addition, the degree of user acceptance will influence the success of the implementation of mixed reality tools as a support element in the execution of a task [59].

C. Occupational Safety and Health

The OSH is a important challenge that affect to the whole people around the world. In 2018 in EU27 3.1 million of the non-fatal accidents occurred, and 3.110 fatal accident, other study, conducted by Hämäläinen et al. reports that in 2014 the world saw 373 million accidents at work [66]. According to the study carried out by Takala, the number of deaths due to professional illness is 2 million and the number caused by an occupational accident is more than 300,000 [67]. The Global Burden of Disease Study of 2015 revealed that 5% of active people's mortality is due to occupational accidents or professional illnesses [68]. Without belittling the importance of protecting the human lives involved, a major component is the economic impact that the safety and health of workers can have on companies and, in general, on the national economy. According to the study by Buerau of Economic Analysis, the estimated cost of work-related accidents and professional illnesses is between $200-550 billions [69]. These data reveal the important problem of workers' safety and health, which justifies the need for tools that help reduce these figures. These circumstances have led to significant increase in OSH research in recent years. Early research in this field comes from other areas, such as medicine [70], although it is already a research field in itself that is of great interest. Its relevance in terms of economic cost and human lives, has led us to analyse the relationships between employees' factors and their working environment that can determine the context for a potential occupational accident [71].

III. Motivation and Methodology

From the study of the OSH situation in Europe and the world we can conclude that we can help the society with new tools and solutions to try to help the amount of people that could have accidents at work. The best way to help the workers is to provide them with the knowledge about how to avoid the accident or, if an accident occurred, how to get a safe him/herself and him/her colleagues. However, it is necessary to help the company with the best tools to train the whole workers team and do this training process like a easy, rapid and cheap way. Today, human augmentation is a technology with great capacity to apply in several user cases. From human augmentation, the authors have selected mixed reality technology as it offers a new way of interaction between the users and digital solution. With mixed reality a user can see a digital 3D object in the same point of view as the real environment, and if we use the HoloLens2 as HMD the user can see and interact with the 3D object thanks to the HoloLens2 gesture and voice recognition. In the literature we find several different approaches of mixed reality with workers to help to do different tasks or help to learn several concepts or processes. So, the researchers create several mixed reality experiences for each paper. Thus, this is the same situation that slows down the application of the virtual reality application in engineering, construction and industry [14] [15] [16]. For this reason and using the literature review the authors introduce the Training-MR. It is a solution to help the mixed reality application in the whole industry 4.0 OSH prevention training process for any type of company. Furthermore, Training-MR helps to democratisation of the technology application because it reduces the cost and time spent by the entities to create, test and publish the mixed reality experiences.

A. Methodology

The creation of the Training-MR was carried out under an agile development method. The methodology chosen by the authors was Scrum. This methodology offers a great capacity to modify the objectives and tasks in a development team depending on the results that occur in each Sprint [72]. Scrum is a development methodology that works very well in research and development projects because the probability of unexpected events is very high as these are projects where uncertainty is important. Thus, development has been divided into two phases.

- Concept phase. The aim of this phase is to reduce the uncertainty of the project. Several proofs of concept (PoC) have been carried out in order to assess whether the mixed reality technology was mature enough to be applied to the project. A laboratory test of the capabilities offered by mixed reality can be seen in Fig. 4. Here the authors check several proofs of concept to use MR as a tool to create MR experiences [73]. During this phase, the state-of-the-art analysis of scientific advances related to the project was also carried out, and the whole requirements list was defined.
- Development phase. This corresponds to the important stage, during which different iterations of the work have been developed in scrum methodology in order to extend the features of Training-MR.

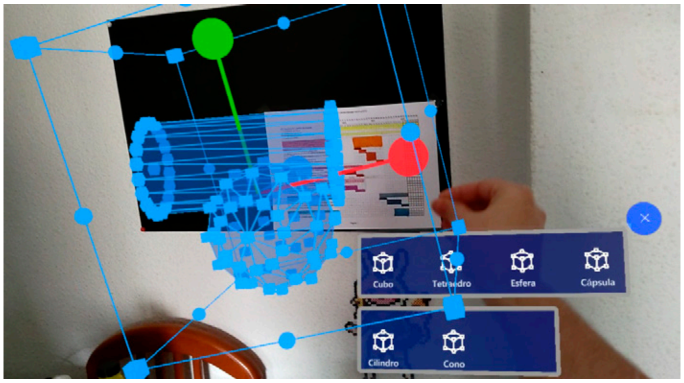


Fig. 4. PoC of the mixed reality capacity to interact with the user and virtual 3D object.

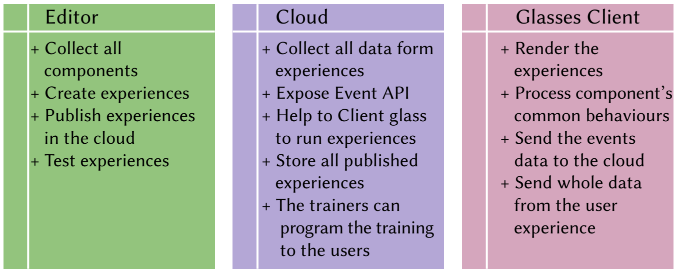
IV. Technical Description

Focusing on the technical description, if we use a high level point of view then the Training-MR has 3 major modules working together. The modules are shown in Fig. 5 and are as follows:

- Editor: the module used to create the whole virtual experience. This module works as a plugin of Unity3D(https://unity.com/) to assist the creation process. Editor is composed of tools to render the scene, create a set of different kinds of components and conFig. them. The Editor is key to solving the cost problem since tools and functions are designed to reduce the time to create a MR experience. At the end of the creation process, the users can send the experience to the cloud for use in the training process.
- Cloud: it is the most important module when the mixed reality

experiences are running. Cloud has the responsibility to manage the whole process in the experience. It is a communications hub that processes all events to determine if it is necessary to run a kind heavyweight algorithm.

- Glasses Client: this module runs in the user devices, and it is the first controller of virtual components behavior. The functions of the Glasses Client are: (1) render the scene, process common behaviour, (2) maintain synchronized event queue with the cloud to process all users actions (3) send all event data to the the Cloud, and (4) collect all data from the user experience to be sent to the cloud.



Editor	Cloud	Glasses Client
+ Collect all components + Create experiences + Publish experiences in the cloud + Test experiences	+ Collect all data form experiences + Expose Event API + Help to Client glass to run experiences + Store all published experiences + The trainers can program the training to the users	+ Render the experiences + Process component's common behaviours + Send the events data to the cloud + Send whole data from the user experience

Fig. 5. Main requirement of the Training-MR modules.

However, the Editor module works isolated at the beginning of the workflow. Two kinds of important data can be highlighted in the workflow of the Training-MR: the virtual experience descriptor and the event descriptor. The first one is a high weight structure of data where any information can be found to create and run the virtual experience. The other one is a lightweight message between glasses and the cloud to process all actions in the MR experience. The virtual experience descriptor is introduced below, together with the event message in the "Communication issue" section.

A. Virtual Experience Descriptor

It corresponds to the core data of any experience in the Training-MR and where information about any element in the virtual experience can be found. This descriptor is an attribute-value file in the JSON language. The most important parts in the descriptor are listing below:

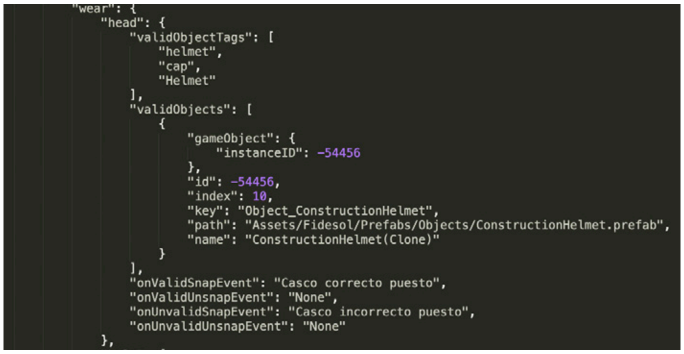
- General data: data to describe the virtual experience, the most important property is the ID Virtual experience, needed to associate the running with the virtual experience in the cloud.
- Scenes descriptions: it constitutes a long list of the components in the catalog. The values of the whole properties of all components can be found here. The exception is the url to download the 3D assets used to render the elements.
- 3D Assets URL: the list of the urls to download 3D assets. These data are split off from the other properties for cybersecurity reasons.

Fig. 6 shows a fragment of the descriptor file. In this example the object "wear" and the parameters such as id, index, onValidSnapEvent, among others, are described.

B. High Level of the Workflow

In normal execution, the Editor does not participate in the run. The reason is that it is usually used to create the experience. Thus, the workflow starts in the glasses client when the user runs our application. The steps to be taken for the execution of the experience are described below and shown in the workflow diagram in the Fig. 7.

1. Start: at the beginning, the users wait for the experience in the hall. The hall is a welcome scene where the users can also interact with some dummy components. These components have been selected to help the user get familiar with the gestures, actions and behaviours from the components.



```
"wear": {
    "head": {
        "validObjectTags": [
            "helmet",
            "cap",
            "Helmet"
        ],
        "validObjects": [
            {
                "gameObject": {
                    "instanceID": -54456
                },
                "id": -54456,
                "index": 10,
                "key": "Object_ConstructionHelmet",
                "path": "Assets/Fidesol/Prefabs/Objects/ConstructionHelmet.prefab",
                "name": "ConstructionHelmet(Clone)"
            }
        ],
        "onValidSnapEvent": "Casco correcto puesto",
        "onValidUnsnapEvent": "None",
        "onUnvalidSnapEvent": "Casco incorrecto puesto",
        "onUnvalidUnsnapEvent": "None"
    },
```

Fig. 6. Example of the mixed reality experience descriptor file.

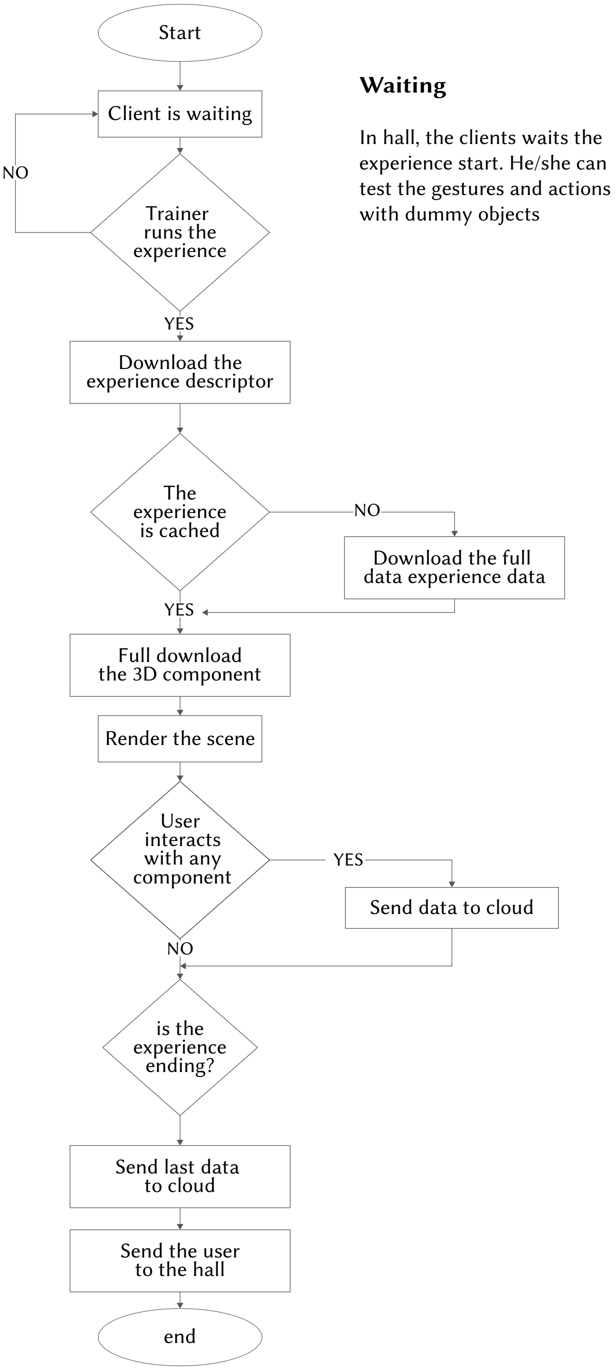

Start
Client is waiting
NO
Trainer runs the experience
YES
Download the experience descriptor
The experience is cached
NO
Download the full data experience data
YES
Full download the 3D component
Render the scene
User interacts with any component
YES
Send data to cloud
NO
is the experience ending?
Send last data to cloud
Send the user to the hall
end
Waiting
In hall, the clients waits the experience start. He/she can test the gestures and actions with dummy objects

Fig. 7. High level of the Training-MR workflow.

2. Load the experience: when the trainer has been selected by the cloud, the glasses client receives one message by web sockets. At this moment, the glasses client checks if the experience exists in the experiences cache then loads from the cache. If it is not cached, the app downloads the full description from the cloud.
3. Start the render process: the 3D engine runs the virtual experience, and prepares all logic subcomponents to be used in the execution, such as: event queue manager, communication abstraction layer, 3D components catalog and many others. Now the workflow is a list of events that have been triggered by the user. The glasses client runs different behaviours from the 3D components and sends several messages to the cloud.
4. End of the experience: this happens when the user successfully performs the last task. The glasses client runs the final actions to close the experience and moves the user to the hall, cleans the memory, sends data from the experience to the cloud, and performs other functions required to prepare the glasses client to run another experience.

C. Training-MR Description

At this point, once the virtual experience descriptor is known, it is needed to describe the Training-MR in more detail. This description is faced by differentiating the following parts: (1st) the cloud and the glasses are described together to facilitate their understanding given their interconnection, (2nd) we introduce the Editor and how it works (3rd) finally, the communication section exposes how the client and cloud share information, and the process by which data are protected from cyberattacks.

1. The Cloud and Glasses Client

It is important to analyse together the Cloud and the Glasses Client, since the design decisions of both are interconnected. Therefore, in a common case of use, a large number of requests are sent to the Cloud from all the glasses clients that are running at the same time. In this way, the microservices architecture in the Cloud is selected Microservices architecture allows the rapid scale of services in peak requests, so the platform automatically reduces these services when the number of requests returns to normal. Services have been designed with Stateless Design Pattern [74], so they are run in an isolated way and endpoints of the public API have been designed to solicit whole data to process. Furthermore, in our solution there are no links or relationships between two services. The technology used to scale services is Docker. On the other hand, the main workflow is the one that takes place in the Glass Client. The Cloud is an unlimited resource to send and request data, from Glass Client's point of view. The Glasses Client workflow acts as an action dispatcher. When an input is detected (for example, user actions), the Glasses Client processes the action according to its code. For this reason the Glasses Client is an Event-driven Architecture solution [75]. The Event-driven Architecture focuses the workflow on event processing. An event could appear for several causes, in our case, the user will be the most important event generator, moreover events from the cloud can also occur. The most important code component in Glasses Client is the Event Queue, where the events are waiting to be processed. Not all events are considered in the same way, so user events are priorities because their delay could cause the freeze or user view error.

The Glasses Client works as a Thin Client in our scheme, so the user could interact with any component that the glasses has renderized. The 3D components present a behaviour similar to the sequence diagram shown in Fig. 8. The sequence starts when the user interacts with the component, which provoques the invocation of the EventManager. The EventManager has the responsibility to start the communication process with the Cloud and invoque de virtual component to modify its properties. Once the response from the Cloud is received by the glasses, virtual component invocation occurs. The properties of the virtual component could be modified in two ways. The first one through a simple action such as launching or moving an object, etc. The second way allows the modification of virtual components with the result of the heavyweight algorithms from the Cloud.

2. Communication Issues

The presented workflow between the Cloud and the Glasses Client must deal with a large number of requests. The Cloud has a Restful API pattern programmed with JSON language. Restful API is a lightweight API standard in Internet services. The Training-MR has a lot of endpoints to manage all information such as User, Student, or Experience, among others although the most important is the ExerciseEvent. The responsibility of the ExerciseEvent is to manage all events for the experiences. However, the API Rest does not handle all communications. When one experience has started, a special message is sent to the Cloud in order to create one web socket between the Cloud and the Glasses Client. The web socket is a channel used by the Cloud to communicate asynchronous data to the Glasses Client. For example, when a heavyweight algorithm process has finished the result should be sent to the Glasses Client that has invoked the algorithm.

Event message An event message is sent to the Cloud caused by a certain trigger or behaviour. The events are usually triggered by the glasses to the Cloud, but there can also be events created and triggered by the Cloud. An example of these events are those from the trainers (actions such as force stop or communication). The Event message has been designed with a short and simple structure.This decision is based on a design that ensures fast message processing. There are different types of messages, but they all have the same properties as the following:

- Type: Type Event.
- ID Virtual Experience: used by the Cloud to identify the virtual

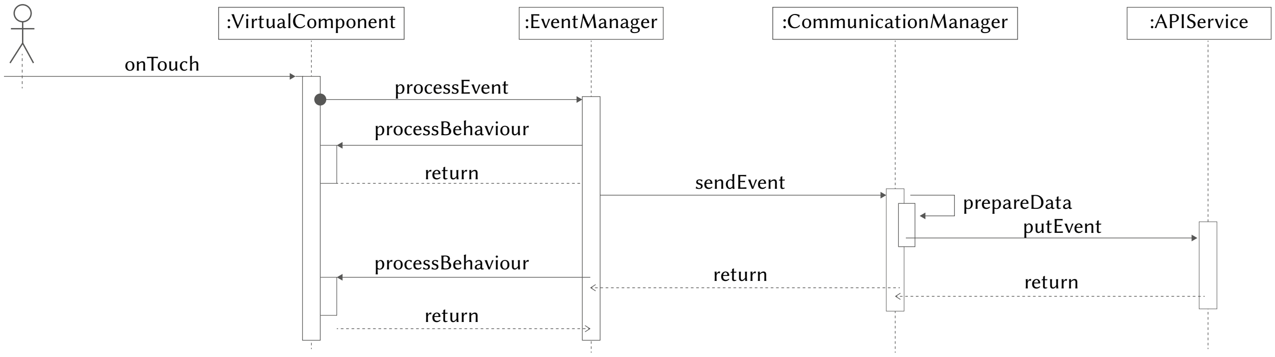


Fig. 8. Sequence diagram of Glasses Client and Cloud.

experience.

- ID running: used by the Cloud to identify a particular execution in any glasses.
- ID user: user running the experience.
- timestamp: timestamp from the glasses.
- time: run time measured in milliseconds. Zero corresponds to the beginning of the MR experience.

Some of the Event messages properties can be customized, which are collected in a list of key-values. Data from the experience are processed in the Cloud to determine whether it is necessary to run a heavyweight algorithm, although all data are always collected and stored in a database. The database selected in the Training-MR is NoSQL instead of the traditional SQL, due to the need for flexibility to store different data sets together and the unknown structure of data from future Event messages. These data will be of particular relevance for analysis in order to know how users interact with mixed reality.

3. Cybersecurity Issues

Cybersecurity is currently a must-have feature in any software or ICT platform, due to the increase in the value of data in recent years. Two types of data coexist in our platform: virtual reality description data, and data related to user interaction in the virtual experience. The first group can contain information from a company, being the end point at which a hacker could steal technical information, for example the end point where a hacker could download a 3D asset and in this way, he/she could steal the technical information. The second group is the data about how the users interact in the experience, this data set could be analysed to determine a lot of information about the industrial process on which the training was designed. To protect these data the solution created has:

- API Key: API requests have parameters about the code that did the request, this is the API KEY, a unique application identifier (code) and it is hardcore in the code so, allows the block of all requests when a cybersecurity attack has occurred.
- Encrypted communication end to end: Communications use SSL encryption for data protection when transmitted over the Internet. In this way, the information will not be understood, if any hacker attempts to sniff the network traffic.
- Encrypted store: Data are encrypted before being stored in the NoSQL database. Moreover, when the glasses caches the data for future use, the virtual experience descriptor is encrypted also.

4. The Editor

The Editor is a module that works isolated from the entire solution. The main objective of the Editor is to provide a toolbox to help content creators. Key components are listed below:

- 3D Viewer. It is the most important requirement because the creator needs to design a 3D scene that will be rendered by the MR glasses. The user selects components from the virtual components catalog to be assigned to the 3D scene. The user must set the properties of these 3D components, such as, physical properties and simple actions (touch, grip, push, etc), among others. At the end of the process, the creator gets the virtual scene with all items positioned on the 3D scene, and all the 3D components completely parameterized.
- 3D Catalog. The 3D Catalog and the 3D Viewer work together. The catalog is a powerful generic 3D component search tool to be used for creating scenes. The 3D components that appear in this catalog are the high abstraction of the tools, situations or triggers.
- •Scene tester. Test scenes are fundamental in the creation process. For example, the creator might need to test the different settings in the scene or probe the relationship between two objects. The Editor allows the creator to test the scenes quickly and easily.
- Publish the mixed reality experience. The Editor and the Cloud are linked. At the end of the process, the creator will upload the mixed reality training to the Cloud. This action will not be available to all users , but only trainers selected by the creator will be able to access this new MR experience.

In order to provide an example of how the Editor works, its application is exposed for selecting the correct electrical wire with an alligator clip. In Fig. 9, a scheme of the proposed virtual scene is displayed. The creator has to describe: the alligator clip, the electrical wires and the triggers (one right option and two wrong). It is important to highlight that these items should be selected by the creator from the Catalog: the physical object for the electrical wires, a hand object to abstract the alligator clip that will be grabbed by the user, and the trigger zone attached to the electrical wires (represented in the scheme by the blue items around the electrical wires). The electrical wires and the alligator clip are physical objects, so the creator has to set their physical properties such as weight, and 3D assets, etc. The triggers are components that do not have 3D assets, weight or physical behaviour, but require a special event when the area is touched by the alligator clip. This contact does not cause movements but the test reaction, to identify whether the action has been successful or not.

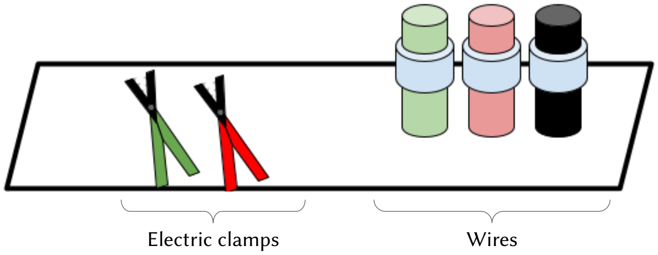


Fig. 9. Example schema of the MR virtual scene example.

Finally, a simple example is provided to describe the workflow of the Editor, although it allows the creator to compose complex behaviours and trigger hierarchical events. Fig. 10 shows the definition of the workflow from a real virtual scene in which the trigger identifies as a failure the lack of worker protection equipment in a welding training.

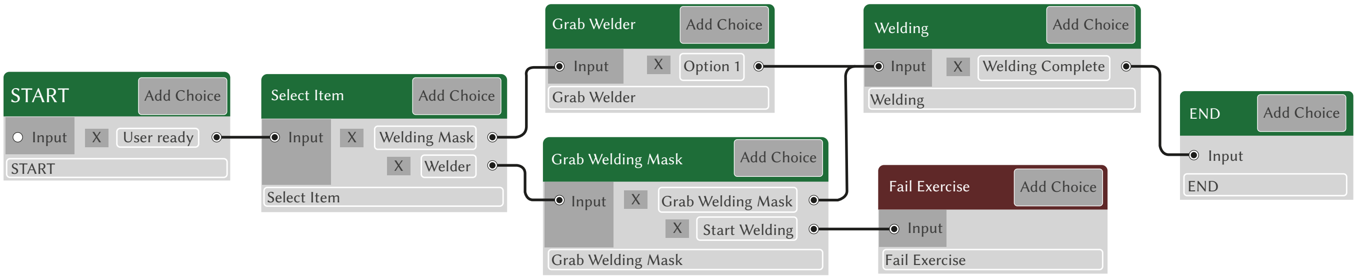


Fig. 10. Example the trigger and behaviour composition in the Editor.

TABLE I. Resume of the MR/VR Solutions

Feature	Training - MR	Blufamsterdam technology	Neurodigital tech	ClassVR
Mixed reality technology	OK	OK	NO	NO
Creation content toolkit	OK	NO	With SDK	NO
Publication capacity	OK	NO	NO	NO
Cyber security design	OK	Without details	Without details	OK

Therefore, the great potential of the Editor makes it possible for users to quickly and easily create mixed reality experiences. Moreover, it is necessary to improve the speed to create MR experience because the economical cost is a high stopper to apply MR as a tool, thus the Training-MR helps to improve training especially in occupational health and safety by not compromising the integrity of the workers.

D. Discussion and Conclusion

Virtual reality and mixed reality correspond to new technologies that can be adopted in the productive processes of companies belonging to different sectors of activity. Researchers have tried to develop solutions that apply augmented reality to solve problems in different fields such as education, industry, engineering, and so on. Mixed reality allows us to go one step further. MR is a powerful tool to improve user training and education as they can interact with both digital and real components at the same time. However, the adoption of MR as a training tool presents the same problems identified for the application of virtual reality [14] [15] [16]. These are the high costs and excessive time required for implementation. The solution to create, test, and publish MR experiences presented in this paper. Training-MR is a solution to all of the problems mentioned above. The main advantages achieved with this tool are listed below:

- From the user perspective, the MR experience is a service from the Cloud. The Glasses Client is a dummy application that only detects the user actions to send requests to the Cloud and process the response data.
- The architecture has been designed to be scalable based on the number of the requests. Therefore, the Cloud can respond to any number of Glasses Clients worldwide.
- The Glasses Client has been designed to prioritize the reaction of user actions to avoid the freezing effect. The Event-driven Architecture enables the prioritization of user events, queuing the least relevant events.
- Training-MR has been built on the principles of cybersecurity and best practices applied to information security. Cybersecurity requirements have been addressed integrated into the platform considering their efficiency and optimization, resulting in a fully secure platform.
- Training-MR, with its Editor, allows to the user to create easy, quickly and quality mixed reality experiences and test these to fix problems o improve the scenes
- A user can share with others the MR experiences created by Training-MR. Training-MR offers the capability to publish the MR experience in the Cloud, and these experiences can be used by other users. In this way, it helps to the democratization of the MR
- The MR experiences have been loaded into the Glasses Client in a quick and easy way. Users can interact with MR experience and improve their knowledge of occupational safety and health.

We compare Training-MR with other public solutions to train the OSH prevention process. Nowadays, the most common situation is that the entities with know-how of the OSH training or experience in virtual reality solutions do not have mixed reality solutions. The most important features to help the easy and rapid application of the mixed reality solutions are used to compare the solutions. These features are: capacity to help to create experiences, mixed reality experiences, experience public capability, cybersecurity design. The Table I resumes our review with other products. It is not common to find solutions with MR technology and usually are for entertainment business. Virtual reality for education and training proposals is more common. Other products are only a set of VR experiences of several kinds of topics. This is the ClassVR case. Rarely find solutions to create content with tools from VR technology owner, and the tools are a SDK to create content by code develop. The public information about the solutions does not explain any cyber security issues.

Our platform is currently in the laboratory testing phase. This phase has taken place after the research team has conducted certain relevant tests, from which a set of MR experiences have been created and will be tested by control users. These users will select an experience, choosing between virtual reality or augmented reality, depending on their knowledge of each technology. After this test with the control users, a final phase called "pilot experience" will be created. In this phase the platform will be tested for training common users in a set of mixed reality controllers. Despite the platform's advantages, this study is not without limitations:

- Gestures. Hololens2 is the most powerful MR device but has a limitation in terms of gesture recognition. It is not possible to recognize a movement that happens behind the user, just as it does not detect an object that is covered by another.
- Multiplayer experience. The platform currently does not allow the creation of a multiplayer MR experience. In the future, special attention will be given to collaborative training.
- Teacher assistance. This has close relationships with the multiplayer. The platform does not currently allow to introduce the teacher's actions into the MR experience as inany learning process in which the teacher interacts with the student.
- Real test. Once the pilot experience is completed, it would be of great importance to test the platform in an industrial environment. To test the components, behaviours, and results of the training in occupational safety and health to validate the MR as a powerful tool for this learning.
- Data analysis. Our platform generates a huge set of data for every MR experience. These data are of great value because their analysis will allow to establish relationships between variables specific to workers with the conditions of the workplace and the use of tools, in order to determine the situations that could lead to an occupational accident or a professional illness.

Acknowledgment

The whole Fidesol team for the great work and capacity in "Interacción y Programación Inteligente para la Industria 4.0 a través de la Realidad Mixta y Fabricación Aditiva" project with code "PY18-RE-0024".

References

[1] A. P. Botha, “Rapidly arriving futures: Future readiness for industry 4.0,” *South African Journal of Industrial Engineering*, vol. 29, nov 2018, doi: 10.7166/29-3-2056.

[2] J. A. Saucedo-Martínez, M. Pérez-Lara, J. A. Marmolejo-Saucedo, T. E. Salais-Fierro, P. Vasant, “Industry 4.0 framework for management and operations: a review,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, pp. 789–801, jun 2017, doi: 10.1007/s12652-017-0533-1.

[3] J. Sengupta, S. Ruj, S. D. Bit, “A secure fog- based architecture for industrial internet of things and industry 4.0,” *IEEE Transactions on Industrial Informatics*, vol. 17, pp. 2316–2324, apr 2021, doi: 10.1109/tii.2020.2998105.

[4] J. L. Ruiz-Real, J. Uribe-Toril, J. A. Torres, J. D. Pablo, “Artificial Intelligence in Business and Economics Research: Trends and Future,” *Journal of Business Economics and Management*, vol. 22, pp. 98–117, oct 2020, doi: 10.3846/jbem.2020.13641.

[5] P. Milgram, F. Kishino, “A taxonomy of mixed reality visual displays,” *IEICE Trans. Information Systems*, no. 12, pp. 1321–1329, 1994.

[6] M.-D. González-Zamar, E. Abad-Segura, “Implications of virtual reality in arts education: Research analysis in the context of higher education,” *Education Sciences*, vol. 10, p. 225, aug 2020, doi: 10.3390/educsci10090225.

[7] J. Wolfartsberger, “Analyzing the potential of virtual reality for engineering design review,” *Automation in Construction*, vol. 104, pp. 27–37, aug 2019, doi: 10.1016/j.autcon.2019.03.018.

[8] J. M. Lombardo, M. A. Lopez, M. López, M. León, F. Miron, J. Arambarri, D. Álvarez, “MOBEEZE. natural interaction technologies, virtual reality and artificial intelligence for gait disorders analysis and rehabilitation in patients with parkinson’s disease,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 6, p. 54, 2019, doi: 10.9781/ijimai.2019.07.003.

[9] Y.-C. Du, S.-C. Fan, L.-C. Yang, “The impact of multi-person virtual reality competitive learning on anatomy education: a randomized controlled study,” *BMC Medical Education*, vol. 20, oct 2020, doi: 10.1186/s12909-020-02155-9.

[10] D. V. Joao, P. Z. Lodetti, A. B. dos Santos, M. A. I. Martins, S. de Francisci, J. F. B. Almeida, “Augmented reality application to assist in on-field activities on the electrical sector,” in *2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, feb 2021, IEEE.

[11] Y. Dan, Z. Shen, Y. Zhu, L. Huang, “Using mixed reality (MR) to improve on-site design experience in community planning,” *Applied Sciences*, vol. 11, p. 3071, mar 2021, doi: 10.3390/app11073071.

[12] E. Bottani, F. Longo, L. Nicoletti, A. Padovano, G. P. C. Tancredi, L. Tebaldi, M. Vetrano, G. Vignali, “Wearable and interactive mixed reality solutions for fault diagnosis and assistance in manufacturing systems: Implementation and testing in an aseptic bottling line,” *Computers in Industry*, vol. 128, p. 103429, jun 2021, doi: 10.1016/j.compind.2021.103429.

[13] “Employer-reported workplace injuries and illnesses 2016,” Bureau of labor statistics. U.S Departament of labor, 2016.

[14] J. M. D. Delgado, L. Oyedele, P. Demian, T. Beach, “A research agenda for augmented and virtual reality in architecture, engineering and construction,” *Advanced Engineering Informatics*, vol. 45, p. 101122, aug 2020, doi: 10.1016/j.aei.2020.101122.

[15] J. Wan, Y. Zheng, Y. Li, H. Mei, L. Lin, L. Kuang, “Oil depot safety inspection and emergency training system based on virtual reality technology,” *IOP Conference Series: Materials Science and Engineering*, vol. 782, p. 042018, apr 2020, doi: 10.1088/1757- 899x/782/4/042018.

[16] J. S. D. Orlean G. Dela Cruz, “Virtual reality (vr): A review on its application in construction safety,” *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 11, pp. 3379–3393, 2021.

[17] R. R. Nick Bostrom, “Ethical issues in human enhancement,” *New Waves in Applied Ethics*, pp. 120– 152, 2008.

[18] T. Garcia, R. Sandler, “Enhancing justice?,” *NanoEthics*, vol. 2, pp. 277–287, nov 2008, doi: 10.1007/s11569-008- 0048-5.

[19] Z. Li, “Ethical problems concerning human augmentation technology and its future aspects,” in *Proceedings of the 7th International Conference on Humanities and Social Science Research (ICHSSR 2021)*, 2021, Atlantis Press.

[20] R. Raisamo, I. Rakkolainen, P. Majaranta, K. Salminen, J. Rantala, A. Farooq, “Human augmentation: Past, present and future,” *International Journal of Human- Computer Studies*, vol. 131, pp. 131–143, nov 2019, doi: 10.1016/j.ijhcs.2019.05.008.

[21] E. Pietrafesa, S. Iavicoli, A. Martini, R. Simeone, Polimeni, “Occupational safety and health education and training: an innovative format and experience,” in *6th International Conference on Higher Education Advances (HEAd’20)*, jun 2020, Universitat Politècnica de València.

[22] P.-E. Boileau, “Sustainability and prevention in occupational health and safety,” *Industrial Health*, vol. 54, no. 4, pp. 293–295, 2016, doi: 10.2486/indhealth.54-293.

[23] A. Simeone, A. Caggiano, L. Boun, R. Grant, “Cloud- based platform for intelligent healthcare monitoring and risk prevention in hazardous manufacturing contexts,” *Procedia CIRP*, vol. 99, pp. 50–56, 2021, doi: 10.1016/j.procir.2021.03.009.

[24] R. Suganya, S. Gowtham, “Individual health and safety monitoring of workers in deep underground mines using IOT,” *Journal of Physics: Conference Series*, vol. 1717, p. 012044, jan 2021, doi: 10.1088/1742-6596/1717/1/012044.

[25] M. B. V. kumar, M. B. Jayasree, M. D. Kiruthika, “Iot based underground coalmine safety system,” *Journal of Physics: Conference Series*, vol. 1717, p. 012030, jan 2021, doi: 10.1088/1742-6596/1717/1/012030.

[26] V. Jayasree, M. N. Kumari, “IOT based smart helmet for construction workers,” in *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, jul 2020, IEEE.

[27] I. Campero-Jurado, S. Márquez-Sánchez, J. Quintanar- Gómez, S. Rodríguez, J. M. Corchado, “Smart helmet 5.0 for industrial internet of things using artificial intelligence,” *Sensors*, vol. 20, p. 6241, nov 2020, doi: 10.3390/s20216241.

[28] M. A. Gigante, “Virtual reality: Definitions, history and applications,” in *Virtual Reality Systems*, Elsevier, 1993, pp. 3–14, doi: 10.1016/b978-0-12-227748-1.50009-3.

[29] S. S. Kardong-Edgren, S. L. Farra, G. Alinier, H. M. Young, “A call to unify definitions of virtual reality,” *Clinical Simulation in Nursing*, vol. 31, pp. 28–34, jun 2019, doi: 10.1016/j.ecns.2019.02.006.

[30] F. Quigley, A. Moorhead, R. Bond, H. Zheng, T. McAloon, “A virtual reality training tool to improve weight-related communication across healthcare settings,” in *Proceedings of the 31st European Conference on Cognitive Ergonomics*, sep 2019, ACM.

[31] E. A.-L. Lee, K. W. Wong, C. C. Fung, “How does desktop virtual reality enhance learning outcomes? a structural equation modeling approach,” *Computers & Education*, vol. 55, pp. 1424–1442, dec 2010, doi: 10.1016/j.compedu.2010.06.006.

[32] C. A. Cohen, M. Hegarty, “Visualizing cross sections: Training spatial thinking using interactive animations and virtual objects,” *Learning and Individual Differences*, vol. 33, pp. 63–71, jul 2014, doi: 10.1016/j.lindif.2014.04.002.

[33] N. Bouali, E. Nygren, S. S. Oyelere, J. Suhonen, V. Cavalli-Sforza, “Imikode,” in *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, nov 2019, ACM.

[34] S. Greenwald, A. Kulik, A. Kunert, S. Beck, B. Frohlich, S. Cobb et al., *Technology and applications for collaborative learning in virtual reality.* 2017.

[35] T. Civelek, E. Ucar, H. Ustunel, M. K. Aydın, “Effects of a haptic augmented simulation on k-12 students’ achievement and their attitudes towards physics,” *EURASIA Journal of Mathematics, Science and Technology Education*, vol. 10, dec 2014, doi: 10.12973/eurasia.2014.1122a.

[36] C. Porter, J. Smith, E. Stagar, A. Simmons, M. Nieberding, C. Orban, J. Brown, A. Ayers, “Using virtual reality in electrostatics instruction: The impact of training,” *Physical Review Physics Education Research*, vol. 16, sep 2020, doi: 10.1103/physrevphyseducres.16.020119.

[37] R. H., “Development of virtual reality training for fire safety education,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, pp. 5906–5912, aug 2020, doi: 10.30534/ijatcse/2020/253942020.

[38] T. Pitana, H. Prastowo, A. P. Mahdali, “The development of fire safety appliances inspection training using virtual reality (VR) technology,” *IOP Conference Series: Earth and Environmental Science*, vol. 557, p. 012064, sep 2020, doi: 10.1088/1755- 1315/557/1/012064.

[39] M. Li, Z. Sun, Z. Jiang, Z. Tan, J. Chen, “A virtual reality platform for safety training in coal mines with AI and cloud computing,” *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 1–7, oct 2020, doi:

10.1155/2020/6243085.

[40] J. M. Lombardo, M. A. Lopez, V. García, M. López, R. Cañadas, S. Velasco, M. León, "PRACTICA. a virtual reality platform for specialized training oriented to improve the productivity," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 4, p. 94, 2019, doi: 10.9781/ijimai.2018.04.007.

[41] M. A. Lopez, J. M. Lombardo, R. González-Crespo, "Educon 2021-creame: human augmentation platform for thecreation of training in educational lakes inherent todangerous situations." 2021.

[42] M. Speicher, B. D. Hall, M. Nebeling, "What is mixed reality?," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, may 2019, ACM.

[43] O. M. Tepper, H. L. Rudy, A. Lefkowitz, K. A. Weimer, S. M. Marks, C. S. Stern, E. S. Garfein, "Mixed reality with HoloLens," *Plastic and Reconstructive Surgery*, vol. 140, pp. 1066–1070, nov 2017, doi: 10.1097/prs.0000000000003802.

[44] Microsoft, "Hololens 2 techspec," [Online]. Available: https://www.microsoft.com/en-us/d/hololens-2/91pnzzznzwcp?activetab=pivot:techspecstab.

[45] Microsoft, "Hololens 2 gestures for authoring and navigating in dynamics 365 guides," [Online]. Available: https://docs.microsoft.com/en- us/dynamics365/mixed-reality/guides/authoring- gestures-hl2.

[46] S. Rokhsaritalemi, A. Sadeghi-Niaraki, S.-M. Choi, "A review on mixed reality: Current trends, challenges and prospects," *Applied Sciences*, vol. 10, p. 636, jan 2020, doi: 10.3390/app10020636.

[47] R. G. Boboc, F. Gîrbacia, E. V. Butilă, "The application of augmented reality in the automotive industry: A systematic literature review," *Applied Sciences*, vol. 10, p. 4259, jun 2020, doi: 10.3390/app10124259.

[48] W. Kurschl, S. Pimminger, J. Schönböck, M. Augstein, J. Altmann, "Using mixed reality in intralogistics - are we ready yet?," *Procedia Computer Science*, vol. 180, pp. 132–141, 2021, doi: 10.1016/j.procs.2021.01.136.

[49] A. Kaluza, M. Juraschek, L. Büth, F. Cerdas, C. Herrmann, "Implementing mixed reality in automotive life cycle engineering: A visual analytics based approach," *Procedia CIRP*, vol. 80, pp. 717–722, 2019, doi: 10.1016/j.procir.2019.01.078.

[50] B. Bejczy, R. Bozyil, E. Vaičekauskas, S. B. K. Petersen, S. Bøgh, S. S. Hjorth, E. B. Hansen, "Mixed reality interface for improving mobile manipulator teleoperation in contamination critical applications," *Procedia Manufacturing*, vol. 51, pp. 620–626, 2020, doi: 10.1016/j.promfg.2020.10.087.

[51] R. Zhang, X. Liu, J. Shuai, L. Zheng, "Collaborative robot and mixed reality assisted microgravity assembly for large space mechanism," *Procedia Manufacturing*, vol. 51, pp. 38–45, 2020, doi: 10.1016/j.promfg.2020.10.007.

[52] A. Siyaev, G.-S. Jo, "Towards aircraft maintenance metaverse using speech interactions with virtual objects in mixed reality," *Sensors*, vol. 21, p. 2066, mar 2021, doi: 10.3390/s21062066.

[53] X. Wang, "Editorial visualization in engineering," *Visualization in Engineering*, vol. 2, mar 2014, doi: 10.1186/2213-7459-2-1.

[54] H. Silva, R. Resende, M. Breternitz, "Mixed reality application to support infrastructure maintenance," in *2018 International Young Engineers Forum (YEF-ECE)*, may 2018, IEEE.

[55] H. Eschen, T. Kötter, R. Rodeck, M. Harnisch, T. Schüppstuhl, "Augmented and virtual reality for inspection and maintenance processes in the aviation industry," *Procedia Manufacturing*, vol. 19, pp. 156–163, 2018, doi: 10.1016/j.promfg.2018.01.022.

[56] A. Fonnet, N. Alves, N. Sousa, M. Guevara, L. Magalhaes, "Heritage BIM integration with mixed reality for building preventive maintenance," in *2017 24º Encontro Português de Computação Gráfica e Interação (EPCGI)*, oct 2017, IEEE.

[57] J. Christian, H. Krieger, A. Holzinger, R. Behringer, "Virtual and mixed reality interfaces for e- training: Examples of applications in light aircraft maintenance," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2007, pp. 520–529.

[58] F. D. Pace, F. Manuri, A. Sanna, D. Zappia, "A comparison between two different approaches for a collaborative mixed-virtual environment in industrial maintenance," *Frontiers in Robotics and AI*, vol. 6, mar 2019, doi: 10.3389/frobt.2019.00018.

[59] S. R. Sorko, C. Trattner, J. Komar, "Implementing AR/MR – learning factories as protected learning space to rise the acceptance for mixed and augmented reality devices in production," *Procedia Manufacturing*, vol. 45, pp. 367–372, 2020, doi: 10.1016/j.promfg.2020.04.037.

[60] S. Lang, M. S. S. D. Kota, D. Weigert, F. Behrendt, "Mixed reality in production and logistics: Discussing the application potentials of microsoft HoloLensTM," *Procedia Computer Science*, vol. 149, pp. 118–129, 2019, doi: 10.1016/j.procs.2019.01.115.

[61] L. Wunder, N. A. G. Gomez, J. E. Gonzalez, G. Mitzova- Vladinov, M. Cacchione, J. Mato, C. L. Foronda, J. A. Groom, "Fire in the operating room: Use of mixed reality simulation with nurse anesthesia students," *Informatics*, vol. 7, p. 40, sep 2020, doi: 10.3390/informatics7040040.

[62] H. F. Moore, M. Gheisari, "A review of virtual and mixed reality applications in construction safety literature," *Safety*, vol. 5, p. 51, aug 2019, doi: 10.3390/safety5030051.

[63] M. Czarski, Y. T. Ng, M. Vogt, M. Juraschek, B. Thiede, P. S. Tan, S. Thiede, C. Herrmann, "A mixed reality application for studying the improvement of HVAC systems in learning factories," *Procedia Manufacturing*, vol. 45, pp. 373–378, 2020, doi: 10.1016/j.promfg.2020.04.039.

[64] A. Rojo, L. Raya, A. Sanchez, "A novel mixed reality solution based on learning environment for sutures in minor surgery," *Applied Sciences*, vol. 11, p. 2335, mar 2021, doi: 10.3390/app11052335.

[65] K. Kounlaxay, S. K. Kim, "Design of learning media in mixed reality for lao education," *Computers, Materials & Continua*, vol. 64, no. 1, pp. 161–180, 2020, doi: 10.32604/cmc.2020.09930.

[66] J. T. T. B. K. Päivi Hämäläinen, "Global estimates of occupational accidentsand workrelated illnesses 2017," in *World Congress on Safety and Health at Work 2017, 3-4 September 2017 Singapore*, 2017.

[67] J. Takala, P. Hämäläinen, K. L. Saarela, L. Y. Yun, K. Manickam, T. W. Jin, P. Heng, C. Tjong, L. G. Kheng, S. Lim, G. S. Lin, "Global estimates of the burden of injury and illness at work in 2012," *Journal of Occupational and Environmental Hygiene*, vol. 11, pp. 326–337, apr 2014, doi: 10.1080/15459624.2013.863131.

[68] H. Wang, M. Naghavi, C. Allen, R. M. Barber, "Global, regional, and national life expectancy, all- cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015," *Lancet*, vol. 388, no. 10053, pp. 1459–1544., 2016.

[69] "Employer-reported workplace injury and illness - 2016," Bureau of Labor Statistics, 2017. [Online]. Available: https://www.bls.gov/news.release/pdf/ osh.pdf.

[70] D. Fan, C. J. Zhu, A. R. Timming, Y. Su, X. Huang, Y. Lu, "Using the past to map out the future of occupational health and safety research: where do we go from here?," *The International Journal of Human Resource Management*, vol. 31, pp. 90–127, sep 2019, doi: 10.1080/09585192.2019.1657167.

[71] A. Kramer, S. Cho, R. S. Gajendran, "12-year longitudinal study linking within-person changes in work and family transitions and workplace injury risk," *Journal of Safety Research*, vol. 75, pp. 140–149, dec 2020, doi: 10.1016/j.jsr.2020.08.009.

[72] K. Schwaber, "SCRUM development process," in *Business Object Design and Implementation*, Springer London, 1997, pp. 117–134.

[73] M. A. Lopez, M. D. Ruiz, D. Alvarez, "Designme-mr: Toolbox for the creation of learning scenes to training in occupation risk prevention with mixed reality." June 2021.

[74] L.-P. T. Kamalmeet Singh, Adrian Ianculescu, *Design Patterns and Best Practices in Java*. 2018.

[75] M. Richards, *Software Architecture Patterns*. 2015.

Miguel Angel López

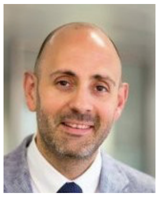
He was graduated in Technical Engineering in Computer Systems from University of Almería, and in Computer Engineering and Master's Degree in Soft Computing and Intelligent Systems from University of Granada. At the moment, he is the CTO at Fidesol where he performs different roles.He is currently a PhD student at International University of La Rioja (UNIR). His research focuses on distributed systems, management, integration and analysis of data, robotics, fuzzy logic systems, human augmented, and the development of virtual/mixed reality environments.

Sara Terrón

PhD in Business and Economics Studies from University of Granada, was graduated in Building Engineering and Master's Degree in Integral Safety in Building from University of Seville and University of Granada. Senior technician in occupational risk prevention. Author of several papers, at Fidesol she currently focuses her research on the technological area.

Juan Manuel Lombardo

PhD in Computer Science from the Pontifical University of Salamanca, was graduated in Economics and Business Administration in the University of Granada, Spain, Diploma of Advanced Studies (DEA) in Economics from UNED, Research Sufficiency in Business Science from the Complutense University of Madrid and Diploma of Advanced Studies (DEA) in Sociology from the Pontifical University of Salamanca. He is CEO at Fidesol and Professor at Andalusia Business School. Dr. Lombardo is the author of numerous articles and research papers published in journals and books of national and international conferences. Visiting Professor at the Private Technical University of Loja (UTPL Ecuador), The National University of the Northeast (Argentina), University Francisco José de Caldas (Colombia), Catholic University of Colombia, Catholic University of Ibarra (Ecuador), University of Lisbon (Portugal) and National Engineering University (Peru). Member of the Knowledge Management committee of AEC (Spanish Association for Quality) and the Institute CICTES (Ibero-American Centre on Science, Technology and Society).



Rubén Gonzalez Crespo

Dr Rubén González Crespo is a full professor in Computer Science and Artificial Intelligence. Currently he is Vice-Rector of Academic Affairs and Teaching from UNIR. He is EiC of the International Journal of Interactive Multimedia and Artificial Intelligence (SCIE), and associate editor in several indexed journals. His main research areas are Artificial Intelligence, Accessibility and TEL. He is advisory board member for the Ministry of Education in Colombia and Spain.

Extensive Classification of Visual Art Paintings for Enhancing Education System using Hybrid SVM-ANN with Sparse Metric Learning based on Kernel Regression

Fei Xu[1]*, Tong Wu[2], Shali Huang[1], Kuntong Han[1], Wenwen Lin[1], Shizhong Wu[1]*, Sivaparthipan CB[3], Dinesh Jackson Samuel R[4]

[1] Academy of Arts & Design, Tsinghua University, Beijing 100084 (China)
[2] Department of Art Design, China Academy of Art, Hangzhou 310002 (China)
[3] Adhiyamaan College of Engineering, Hosur (India)
[4] Oxford Brookes University, Oxford (United Kingdom)

Received 28 March 2020 | Accepted 2 April 2021 | Published 1 October 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

In recent decades, the collection of visual art paintings is large, digitized, and available for public uses that are rapidly growing. The development of multi-media systems is needed due to the huge amount of digitized artwork collections for retrieving and archiving this large-scale data. This multimedia system benefits from high-level tasks and has an essential step for measuring the similarity of visual between the artistic items. For modeling the similarities between the artworks or paintings, it is essential to extract useful features of visual paintings and propose the best approach for learning these similarity metrics. The infield of visual arts education, knowing the similarities and features, makes education more attractive by enhancing cognitive development in students. In this paper, the detailed visual features are listed, and the similarity measurement between the paintings is optimized by the Sparse Metric Learning-based Kernel Regression (KR-SML). A classification model is developed using hybrid SVM-ANN for semantic-level understanding to predict painting's genre, artist, and style. Furthermore, the Human-Computer Interaction (HCI) based formulation model is built to analyze the proposed technique. The simulation results show that the proposed model is better in terms of performance than other existing techniques.

Keywords

Support Vector Machine, Human-Computer Interaction, Artificial Neural Network, Sparse Metric Learning, Feature Extraction, Machine Learning, Visual Arts Education, Digitization of Paintings.

DOI: 10.9781/ijimai.2021.10.001

I. Introduction

In this modern era, the fine-art collections have been digitized in a vast amount and can be used by several researchers to study and analyze their importance. These collected data are distinguished as modern, contemporary, and classical artworks. Several multimedia systems are developed for retrieving and archiving these data. In the early modern days, the most common collections are meta-data describing the theme, the artist, the type and data of art curators and historians, etc. [1] [2]. The present artwork is displayed in the online galleries, useful for developing recommendation systems that retrieve similar paintings, which users wish to buy. Moreover, the investigation of visual similarity metrics should be highlighted, which is much needed for optimizing the painting-domain [3].

Digital systems are categorized and recognized the scenes and objects in videos and images by computer vision. The advances of this technology are widely distributed due to the surveillance cameras installed in every place. The enlightened inferences are made when a person views a painting rather than just identifying the chair, the figure of Christ, or tree. Any individual can assume the genre, year of creation, and painting style without knowing visual-art training [4]. The extent of exposure and comprehension of art history by the spectator predicts the precision of their conclusions. Judging and understanding such complex visual ideas is a wonderful opportunity to grasp humanity. The main goal of this research is to expand a semantic-level judgment machine for predicting the painting's genre, style, and artist. This knowledge provides the similarity measurement is optimized in the available domain of art- history simplification. In this research, several researchers benefit from the concept behind the evolution in art analysis based on the computer system. The measurements intimate the art historian's skill to identify painting according to the artist who created it, its theme and genre. Visual characteristics are extracted from the image in the first step. These visual characteristics vary from low to high. In the next step, we learn how to modify these functions for various classification activities by studying the necessary dimensions. The metric has been learned to develop paintings from a big, raw, visual space into a meaningful, much smaller space. The third approach projects visual attributes using various metrics and fuses the resulting optimized spaces to achieve a

* Corresponding author.

E-mail addresses: xufeizhen@mail.tsinghua.edu.cn (Fei Xu), wushiz@vip.sina.com (Shizhong Wu).

final function vector for classification. In this low-dimensional space, a classifier can quickly be applied to large sets. This is an effective technique since each approach uses various parameters to accomplish the measure of the similarity [5].

Moreover, many digitized visual-art datasets are used to evaluate systematic methods [6] comprehensively. The paintings are described in various concepts by the artists in whom stylistic elements like texture, line, space, tone, and color are used. Secondly, principles like unity, pattern, contrast, movement, variety, proportion, and balance are also used. Several visual features are investigated and engineered by the researchers in computerized art analysis. Artistic concepts are encoded like color and brush strokes and low-level features like color histograms and texture statistics. When paintings are digitized, the variations of texture and color are highly vulnerable, also the age of paintings can affect the color [7]. It is inconvenient to design visual features of the concepts mentioned above of art. A deep neural network is recently used to advance computer vision, having a major advantage of feature learning from the given data rather than fully engineering the concept. However, learning visual features is a major concern and impractical, consisting of visual art concepts. Extensive annotation is required for every image's concepts within the huge testing and training dataset [8]. Human-Computer Interaction (HCI) is a computer model that uses the dialogue between the system and man language. It determines the information exchange between the computer processes and people. The identification and segmentation of images play a vital role by HCI, and it is the human visual perception [24]. Enhancement in education takes place by visual art learning and development in cognition. Visual art increases students' thinking capabilities, mental analysis, solving issues, creative thinking, reasoning ability, conceptualization, classifying, and so on. These lead to association with enhanced cognitive development as it has a prominent feature, which consists of temperament and personality.

An alternative strategy is followed in this paper for the challenge mentioned earlier to learning or engineering appropriate visual features of paintings. Various visual elements are investigated mainly that range from low-level to semantic-level. The proposed sparse metric learning based on kernel regression is used for obtaining similarity metrics of various paintings that make effective use for the prediction tasks, namely genre, artist, and style classification. In this paper, many visual features are investigated and a learning methodology is proposed for the above-explained prediction tasks. Here, sparse metric learning-based kernel regression is proposed for optimizing the prediction tasks, and the simulation results are compared with other existing metric learning techniques. Also, a hybrid SVM-ANN model is proposed for improving the classification performance based on the three prediction tasks. The metrics' primary goal is to evaluate the experiment for three separate tasks, including genre, artist, and style prediction. In the following pages, the metric efficiency of the SVM-ANN hybrid classification is analyzed in several features. The proposed Sparse Metric Learning-based Kernel Regression (KR-SML), firstly, as depicted in the figure, extracts visual characteristics from images of paintings. On each of these prediction tasks and a similarity metric tailored is applied, i.e., style-optimized metric, genre-optimized metric, and artist-optimized metric. Any metric induces a projector to a feature space that is optimized for the task. If the metric is mastered, we project the raw visual characteristics into a new optimized function space and SVM-ANN learns the required prediction task. Visual characteristics and metric methods are used to acquire an optimized measure of resemblance between paintings. This provides a computer capable of making semantic decisions relevant to aesthetics, such as an identification of a painting's theme, genre, and artist, and delivering optimized similarity measures based on the knowledge of art history analysis accessible. Our analyses demonstrate the importance of using this indicator of similarity.

The main contribution of the study is:

- Designing KR-SML to obtain detailed visual features of painting and improve accuracy.
- Analyzing the HCI based formulation model to evaluate the proposed technique.
- Once the numerical results have been obtained, the proposed KR-SML uses SVM-ANN for semantic-level understanding to predict painting characteristics and classify paintings, enhancing the genre and artist prediction compared to other methods.

The remaining paper is structured as follows: section I corresponds to the present introduction to the work and sector II describes the related works. In section III, the KR-SML method has been suggested for improving the accuracy of painting and style predictions. In section IV, the numerical results have been obtained. Finally, section V concludes the research paper.

II. Related Works

In this modern technology, computer systems are used for a different set of tasks in paintings. Image processing methods help art historians measure tools such as mathematical brushstroke quantification, pigmentation analysis, etc. Several researchers study the encoding of information about the paintings to find suitable features that could help classify visual art. Major research concerns the painting classification by utilizing low-level features, including shadow, edges, color, and texture. Lombardi [9] analyzed the artist classification's feature types from a small set using unsupervised and supervised machine learning techniques. The proposed methodology identifies the painting's style for finding the artist who designed it.

A comparative study of the style classification task is presented by Arora et al. [10]. Low-level features like Color Scale-invariant feature transform (SIFT) and SIFT are evaluated versus semantic-features like Classemes that include the image object present in it. The authors concluded that the semantic level features perform much better than the low-level for this task. Our study attempts to create a technology that will be able to make semantic judgments on an aesthetic level like predicting the style, genre, and artist of a painting, along with to have optimized similarity actions based on the information available in the field of art historical perception. The role of style classification determines low-level characteristics and the color semantic level characteristics that encode the image object presence. Semantic-level features for this role were found to surpass substantially low-level features. The evaluation of this performance is done in a small dataset that contains 70 paintings with seven styles. Carneiro et al. [11] indicate that color features and a low-level texture have not been reached because they define the visual type of pictures as unpredictable texture and color patterns.

Furthermore, metric learning techniques are used by Saleh et al. [12] for detecting influence paths between the painters, which are based on their paintings. The authors used the HOG feature of low-level and optimized using three metric-learning approaches. Bar et al. [13] identified the style based on characteristics by proposing a convolutional neural network for image categorization. Takeda et al. [14] analyzed the denoising in images by proposing Kernel Regression (KR) algorithm and used local pixel-space statistics for learning the Mahalanobis matrix. Moreover, this proposed method is restricted to particular applications and cannot be presented for all cases.

Karayev et al. [15] analyzed the deep features which have been used for several performances in different fields, also the performance for hand-crafted features such as GIST (Gradient information scales and orientations), color histogram, and visual saliency are efficiently processed. N. Senthil Murugan and G. Usha Devi [16]-[18] analyzed the machine learning concepts and proposed hybrid models for analyzing a large amount of data, and several features have been processed. G Manoharan et al. [19] analyzed the human interaction with computers for analyzing the big data. The intelligent and adaptive model of HCI is discussed by Z Duric et al. [25] and analyzed human motion using compute vision. The author also discussed the models performing low arm movement detection, gaze analysis, and face processing. Cedras and Shah [26] presented the categories within the motion classification extraction based on motion correspondence or optical flow. The problem of human-motion capture is defined as the recognition of action, individual recognition, body estimation, and configuration. Peterson [27] suggested a set of behavioral parameters associated with enhanced cognitive development based upon a review of the brain-mind's science. Gredler and Shields [28] state that visual art and instruction play a crucial role in children's mental improvement. Academic ideas are important in conceptual development and ultimately leading to the development of concepts.

To overcome the existing issues, a KR-SML has been proposed for the similarity measurement of visual features. The proposed MHCBTF used the SVM-ANN method to categorize the measurements and classify the paintings according to style, genre, and artist.

III. Proposed Methodology

In this section, the most suitable combination for the visual features is obtained and proposed KR-SML metrics are used to improve accuracy in similarity measurement. Furthermore, the measurements which categorized the paintings in terms of genre, artist, and style are used for classification based on hybrid SVM-ANN. Firstly, the visual features are extracted from the particular image that ranges from low-level to high-level. Secondly, the proposed metric based on KR-SML is used to adjust the extracted features that are processed for various classification tasks. Based on this metric learning, the paintings can be projected from a high-dimensional to low-dimensional space, which is more meaningful. Finally, the proposed hybrid SVM-ANN is used for classifying the painting based on the low-dimensional features.

A. Dataset Collection

The online dataset named "Wikiart Paintings" [20] is publicly available and contains a large collection of digitized-artworks used for the proposed methodology. The dataset contains 81,499 images of fine-art paintings, and 1199 artists range from the 15th century to the modern artist. Moreover, 27 different styles are included in this painting, like Byzantine, Abstract, Baroque, etc., and 46 various genres such as Landscape, Interior, etc.

Previous researchers used numerous sources to gather minimal data volume regarding genre, and style with restricted heterogeneity. Moreover, the classification is done automatically for the paintings in terms of genre, style, and artist using the visual features extracted by applying computer-vision algorithms. The tasks contained in the existing works have their limitations and challenges. In particular, the date's subset is used for style classification with 27 styles in which each one consists of 1500 paintings and has a total number of 78,439 images. The genre-classification uses the subset of 10 genre-classes containing 1500 paintings, having a total of 63,721 images. For the artist's classification, 23 artists subset is used, which contains 500 paintings with 18,589 images. Table I shows the set of the genre, artist, and style labels.

B. Visual Attributes

The state-of-the-art representative is investigated in this work, which includes two main categories:

- Low-Level Attributes-The GIST features are extracted for capturing the visual information of low-level that is holistic attributes that are developed for scene categorization; Furthermore, the real-valued 512 GIST features are represented, which indirectly captured the image's ruling spatial-structure.
- Semantic-Level Attribute-Here, three images represented based on an object are extracted, such as Picodes, CNN-based variables, and Classeme, for the semantic representation. From these three variables, the image's object-category confidence is represented by each element's feature vector.

TABLE I. List of Genres, Artists, and Styles

Task-Name	List of Members
Genre	Abstract-Painting; Genre Painting; Cityscape; Portrait; Landscape; Still Life; Religious Painting; Sketch and Study
Artist	Boris Kustodiev; Childe Hassam; Edgar Degas; Nicholas Roerich; john Singer Sargent; Marc Chagall; Salvador Dali; MartirosSaryan
Style	Action Painting; High Renaissance; Symbolism; Realism; Rococo; Minimalism; Cubism; Moder Art; Abstract Expressionism; Pop Art; Post Impressionism; Ukiyo-e; Synthetic Cubism; Pointillism; Romanticism

C. Classification Using the Proposed Methodology

The suggested Sparse Metrical Learning based on kernel regression to remove visual features is processed with artist and stylist to identify paintings according to their Genres, defined by a Hybrid SVM-ANN model. The main usage of learning metrics is to find the real-valued function based on pair-wise $F_M(z, z')$ and it is symmetric, non-negative, obeys the inequality triangle, and zero is returned if and only if z and z' are the same-point. The optimization problem for training the function as a general form is given in Eqn. (1):

$$\min_M ls(M, X) + \gamma R(M) \tag{1}$$

Two phases are included in this optimization in which the quantity of loss ls (M, X) for data samples X using metric M and the regularization term R(M), which is adjusted, are used. The accuracy is given by the first term from the metric trained, and capability is estimated from the second term for new data for avoiding the overfitting of the model.

KR-SML: Firstly, the visual features are extracted, as shown in Fig. 1, from the images present in the paintings. The prediction tasks for each image are learned by optimizing the proposed metric model. Each metric induces the projector to a corresponding space for the appropriate mission.

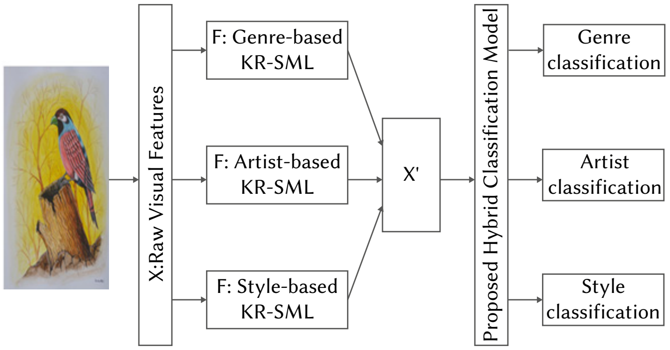


Fig. 1. Proposed Classification Model.

Moreover, the metric matrix of Mahalanobis is trained using the training dataset in which the objective of KR-SML is significant for the proposed work for ensuring the metric matrix is sparse, and error is small. Therefore, the loss function is made for learning the metric matrix of Mahalanobis of Kernel Regression, and the norm regularization with mixed (2, 1) over M is learned to be minimum. The proposed model of KR-SML is represented in Eqn. (2).

$$Ls(M) = \sum_{j=1}^{T_{training}} \left(x_j - \hat{x}_j\right)^2 + \sigma \|M\|_{(2,1)} \tag{2}$$

Where,

$Ls(M)$ is defined as the loss function at a minimum. The norm regularization is non-differentiable and non-convex in the objective function. The proposed algorithm of KR-SML is given below,

```
Step 1: Start
Step 2: Input Adapted step-size β, Matrix M, Adapted step-size σ
        for $Ls(M)$, μ-stop criterion, p ← 0
Step 3: Do p ← p + 1
Compute $GF^{p-}$ objective function at $p^{th}$ iteration
Where, $GF^p = 2\sum_{j=1}^{T_{training}} \left(\hat{x}_j - x_j\right) \frac{\sum_{i=1}^{k} (\hat{x}_j - x_i) Z_{ji} \vec{y}_{ji} \vec{y}_{ji}^{N}}{\sum_{i=1}^{k} Z_{ji}} + \theta I$
$M_{(p)} \leftarrow M_{(p-1)} - \beta GF^p$ is used for updating metric matrix
$M_{(p)} \leftarrow E^N \Delta_+ E$
The objective function value is computed
Until $\left|Ls(M_p) - Ls(M_{p-1})\right| \leq \tau$
Step 4: Output of the M is produced
Step 5: End
```

Using the proposed metric learning (KR-SML), the features' dimensionality can be reduced when the M-metric is in low-rank. More specifically, knowledge is necessary for the ground truth of the input paintings used in a supervised-mode for non-linear and linear cases to research the most wanted metric. The quantity of regularization or the form of M is used for differentiating the various approaches. Moreover, the hybrid SVM-ANN classifier is used to classify metrics based on Style, Artist, and Genre.

D. Cognitive Enhancement

Education in children has an enormous effect on their development. These include emotional, physical, and development in cognition of students. Enhancing the student's cognition refers to improving the thinking capacity of students. It consists of knowledge enhancement, solving problems, skills improvement, and characters. Enhanced cognitive development leads to brain development. Peterson [27] states that the development of languages is the key focus on the development of cognition in students. Cognitive development in neuroscience mainly influences the process of bridging in the educational development of students. Efforts between neuroscience, psychological cognition, and enhanced education have absorbed how people obtain and utilize knowledge. Measuring the test of mental ability of how learning occurs in students is inclined by a grouping of genetic programming, maturation status, and environmental problems. Fig. 2 shows the educational enhancement using visual art, where fine art is used for classification using the proposed methodology. After classification, feature extraction takes process by splitting the texture and shape features separately from the visual art. Feature comparison amongst the shape and texture feature takes part in which the similarity between them is matched. Students visualize the art made after comparing features followed by analysis for making their personality and educational enhancement. The analysis process mainly constitutes the student cognitive development, educational improvement by making learning better by visual learning, and increases students' eye-hand coordination.

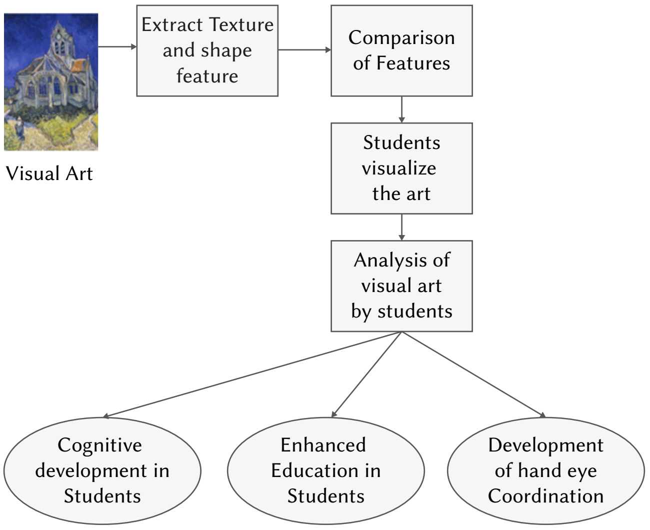

Visual Art
Extract Texture and shape feature
Comparison of Features
Students visualize the art
Analysis of visual art by students
Cognitive development in Students
Enhanced Education in Students
Development of hand eye Coordination

Fig. 2. Enhanced Education in Students using Visual Art.

E. HCI Based Formulation Model

In this section, an interactive system based on a software generation model is used for analyzing the human-interaction for the proposed approach. Two main subsystems are used in this HCI model: Query-Formulation Subsystem and Software-Generation Subsystem. The painting image is formulated by using the query formulation-interface. The generation subsystem of software is queried by formulation-subsystem with objective representation based on the proposed approach.

A software generator builds an appropriate software-based program, and images classified based on the proposed approach are yielded by executing this software based on the test set. The formulation could be reconsidered based on the predicted resulting image, and the new query is submitted with modification in the description of the input. The process gets stopped if the users get the resultant image. For conducting the experiments, the proposed KR-SML is used to extract features in the paintings, which are classified using the hybrid SVM-ANN model and stored as a library, and processed as software, and implemented using HCI architecture. Fig. 3 shows the generation system of the proposed HCI for analyzing the painting image.

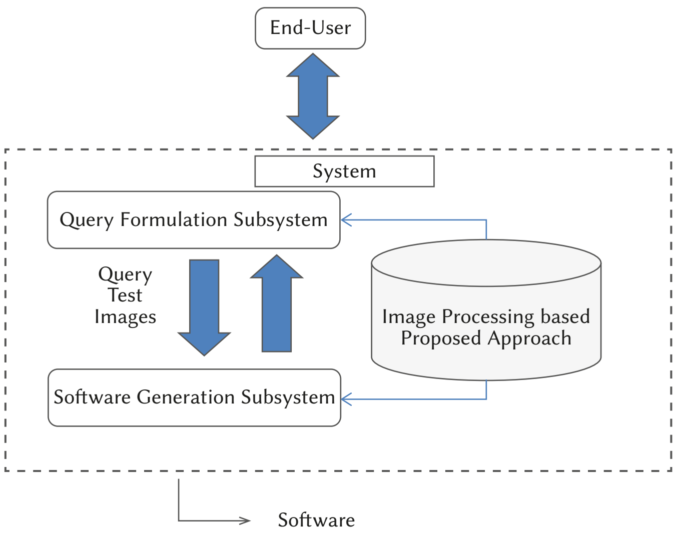

End-User
System
Query Formulation Subsystem
Query Test Images
Image Processing based Proposed Approach
Software Generation Subsystem
Software

Fig. 3. HCI based Architecture using the Proposed Approach.

IV. Simulation Results

A. Simulation Setting

The GIST features are extracted as visual attributes in low-level such as Picodes, Classeme, and CNN-based variables as high-level semantic-features. The implementation of Torralba and Oliva [21] is followed in this research to gather a feature vector of 512 dimensions. The Bergamo et al. [22] implementation is used for Picodes and Classeme, which results in 2045 dimensions for Picodes and 2658 dimensions for Classeme. Furthermore, a 1000 dimensional variable vector is extracted by using Lenc and Vedaldi [23]. The dimensionality of the feature vector is higher than produced images based on object representation and GIST features. The main intention of metric learning is to analyze the experiment labeled for three various tasks of Genre, Artist, and Style Prediction. The metrics performance is investigated in the following sections on various features for classifying using the hybrid SVM-ANN. All metrics are learned from segment 3 for all the 15 styles in the paintings present in a given dataset.

B. Classification Based on Style

Table II shows the results of style classification using the hybrid SVM-ANN after processing various metrics on a set of variables. Columns below contain different characteristics and metrics used before grouping the types in rows for calculating attributes. The ITML and Boost metric approaches give high accuracy for the classification of style for various features. Moreover, the proposed KR-SML approach produces a high accuracy for all types of features when classified using the hybrid SVM-ANN model for almost all the extracted visual features. Fig. 4 shows the overall accuracy of the proposed model.

TABLE II. Style Classification

Metrics/Variables	Classemes	CNN-based	GIST	Picodes	Dimension
Boost-Approach	33.45	16.12	16.07	28.58	512
ITML-Approach	30.67	15.19	13.04	28.42	512
LMNN-Approach	27	16.84	12.53	24.12	100
NCA-Approach	28.8	16.33	13.27	24.68	27
Proposed Approach	**41.34**	**21.34**	**19.87**	**31.34**	**512**

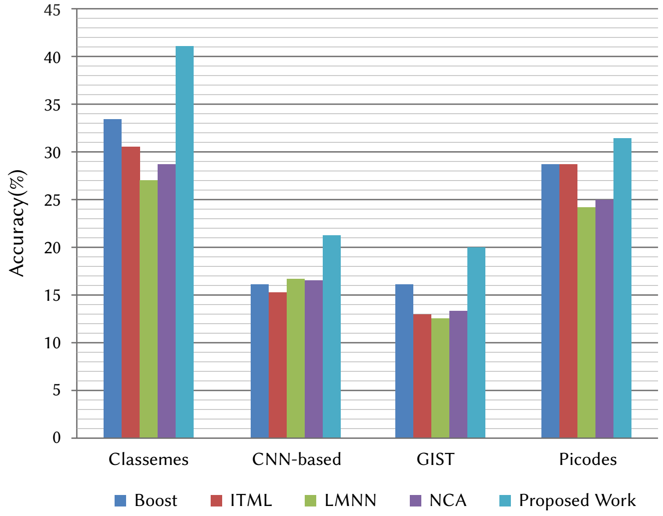


Fig. 4. Accuracy based on Style Classification.

C. Classification Based on Genre

In this classification, a total of 8 genres are used from the dataset for getting several samples reasonably for every task. Table III shows the proposed metric model's classification performance based on eight genres using the hybrid classifier. Various features and metrics which are used for computing the distance are represented in table III. The genre classification performance using the proposed method gives high accuracy compared with other existing classifiers. Moreover, the total number of genre collections is lesser than the style collected in the dataset. Fig. 5 shows the overall accuracy of genre classification based on the proposed approach.

TABLE III. Genre Classification

Metrics/Variables	Classemes	CNN-based	GIST	Picodes	Dimension
Boost-Approach	57.87	57.33	31.02	46.15	512
ITML-Approach	57.88	57.32	33.11	46.83	512
LMNN-Approach	54.98	54.32	39.07	49.97	100
NCA-Approach	51.34	52.76	30.45	49.56	10
Proposed Approach	**62.87**	**63.56**	**43.56**	**51.29**	**512**

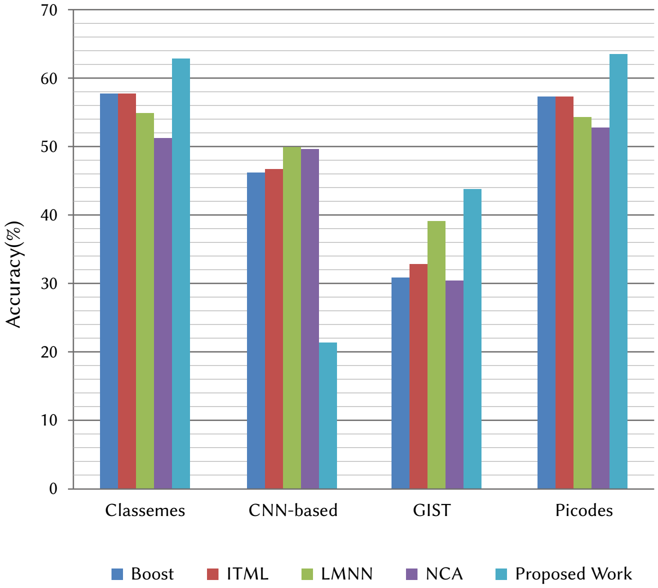


Fig. 5. Classification based on Genre.

D. Classification Based on Artist

Table IV shows the accuracy of the proposed metric approach using the hybrid SVM-ANN approach in terms of features extracted from the images for eight artists and compared the results with other models. Based on the maximum confidence, the artist is determined from the images. The proposed model performance shows a high accuracy when classifying based on an artist compared to other metric learnings. The dimension used for the artist classification is 512, and the accuracy obtained for the features of Classemes, GIST, Picodes, and CNN are higher when comparing with Boost, ITML, and LMNN approaches. Fig. 6 shows the results of the proposed approach.

TABLE IV. Artist Classification

Metrics/Variables	Classemes	CNN-based	GIST	Picodes	Dimension
Boost-Approach	57.72	55.50	25.75	29.56	512
ITML-Approach	51.88	53.87	19.95	31.06	512
LMNN-Approach	53.97	53.98	20.42	30.93	100
NCA-Approach	49.61	19.65	21.77	21.33	23
Proposed Approach	**59.13**	**59.23**	**28.93**	**39.12**	**512**

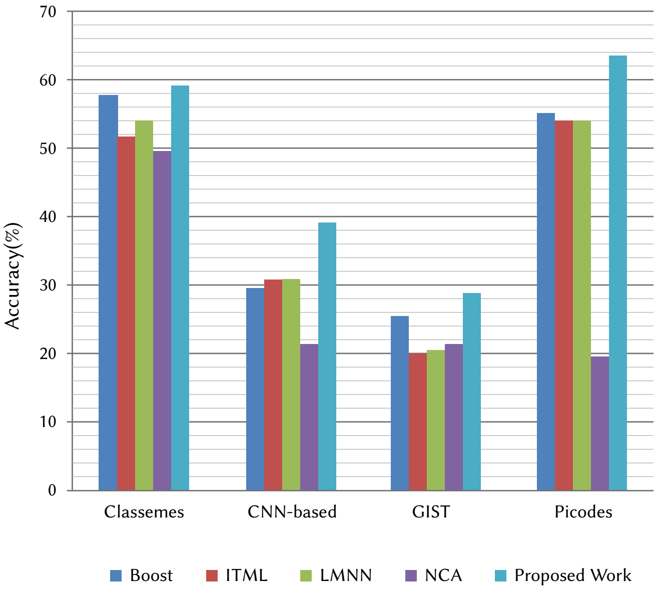


Fig. 6. Classification based on Artist.

E. Metric and Feature Integration

The integration of the metric and features of the proposed model is analyzed to find each approach's performance. The classification performance is analyzed for the task mentioned before by combining the various extracted visual features. Table V shows the results of the integration performance based on the hybrid classifier. The style classification performance is made by half of the images taken from the dataset. The accuracy achieved by the proposed model is about 51.23% when using the features of Picodes, GIST, Classeme, and CNN, while the other approaches like LMNN, NCA, ITML, and Boost produce less accuracy of 45.94%, 40.61%, 45.05%, and 41.74%.

TABLE V. Classification Based on Integration

Model	Artist	Genre	Style
Boost-Approach	61.24	58.51	41.74
ITML-Approach	60.46	60.28	45.05
LMNN-Approach	63.06	58.48	45.97
NCA-Approach	55.83	64.34	40.61
Proposed Approach	68.45	64.34	49.78

Moreover, the compact representation of features is learned in the proposed model by performing the state-of-the-art. The feature vector analyzed from the proposed metric learning is more efficient and useful for representing the images, and the best accuracy for the classification is obtained. The proposed work is considered for image retrieval application in future research. From these results, the proposed approach outperforms the state-of-the-art works and the capacity for image representation is reduced more than 90%. Fig. 7 shows the overall accuracy of the proposed classification performance for Style, Artist, and Genre and compared with other existing approaches.

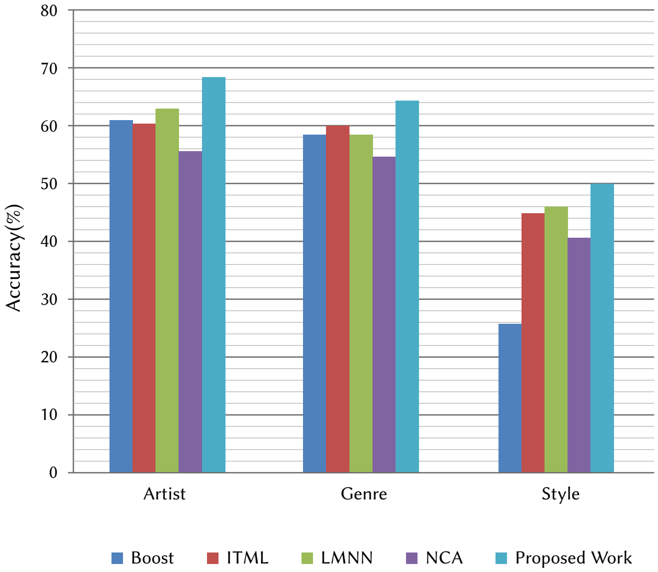


Fig. 7. Classification Performance based on Style, Artist, and Genre.

F. HCI Based Image Identification

Fig. 8 shows the image detection for the given test inputs in which the identification of painting images is analyzed and detected using the interactive software generation system. After the user queries an input image, the generation system analyses and detects the images based on the proposed approach, and the resulting image is displayed. Furthermore, the HCI system reconsiders the user's formulation if he/she submits a new query.

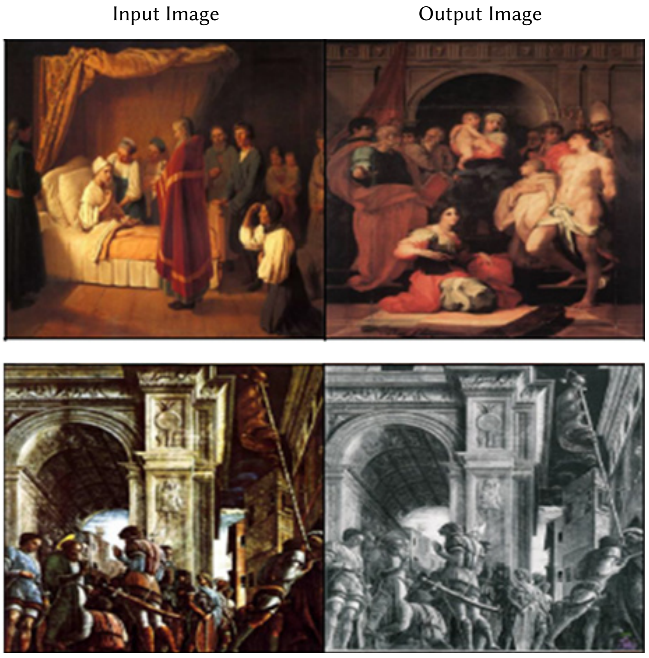


Fig. 8. Resulting Image using the HCI system.

V. Conclusions

In this research, the proposed KR-SML metric learning is investigated for the painting dataset, and various visual features are extracted and its similarity and performance are measured from fine-art-paintings. Several media systems for the recovery and archiving of these data are created. The early modern collections contain meta-data documenting the subject, the artist, the art curators, historians, etc. A wide selection of digitized artworks used in the proposed methodology is available on-line and called Wikiart Painting. The similarity measurement between the paintings is implemented based on the proposed metric learning and classified using the hybrid SVM-ANN model by using three major concepts. Metric learning techniques are implemented to measure the resemblance of various visual characteristics of the fine art paintings series. Meaningful measures are used to measure the resemblance between paintings. The metrics are learned in a supervised way to get paintings from one principle closer to far from each other. Three principals have been used in this work: Style, Genre, and Artist. We used these learned metrics to transform raw visual attributes into a separate space that can enhance the output of three key tasks, including classifications of styles, and genres. To test the efficiency of the above activities, our comparative studies were carried out on the largest publicly accessible data collection of fine-art paintings. These are Genre, Artist, and Style. The accuracy obtained from the proposed model's performance results is about 68.45% for the artist, 64.34% for Genre, 49.78% for style classification, and the other metric approaches. The visual features extracted based on the metric learning are Classemes, GIST, Picodes, and CNN, which classifies the tasks using a hybrid approach. Prediction of the type, artist, and style of painting is created for semantic comprehension. In addition, the formulation model based on metric learning is designed to evaluate the methodology proposed. The findings of the simulation demonstrate that the model suggested is more effective than other strategies. The feature vector size is reduced by more than 90% when using the KR-SML metric learning for classification tasks. The consequence is that we train the SVM classifier on top of dimensional vectors. This outperforms state of the art and offers a clearer depiction of the pictures, which decreases the room by 90%.

Furthermore, the HCI system based on model interaction and formulation is used for obtaining the user's query for gathering paintings image in which implementing the interactive-software generation system. Features extracted from visual art help students visualize the art by analyzing the feature, enhancing cognitive development, improving student curriculum, and maintaining hand-eye coordination. As part of an understanding technique, visualization encourages students to consider the actual scale of unfamiliar objects by contrasting them with familiar objects and using a technique to allow students to create cognitive representations more easily. Therefore, the educational system can be enhanced using visual art by teaching it in the curriculum. As future work, the image retrieval task and recommendation system are appropriate and can be verified by using the proposed approach. Several annotations based metric learning can be analyzed for better feature extraction and classification. Making this visual art-based learning can improve the mental as well as cognitive development in students. The inclusion of these visual art-based learning will make educational systems enhanced. Also, new software generation systems of HCI can be implemented for improving performance.

Acknowledgments

China Postdoctoral International Exchange Program (Grant No. 2019.92).

References

[1] A. E.Abdel-Hakim and A.A.Farag, "CSIFT: A SIFT descriptor with color invariant characteristics". *In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'06) Vol. 2, 2006, June, pp. 1978-1983. IEEE.

[2] R. Arnheim, "A plea for visual thinking," New essays on the psychology of art. University of California Press, 1986, pp. 135-152.

[3] T. P. Beebe, Z. Voras, K. de Gheraldi and J. Mass, "Surface Analysis of Fine Art Paintings: Studying Degradation Mechanisms with a Systematic Approach," *Microscopy and Microanalysis*, vol. 24, no. S1, 2018, pp. 1050-1053.

[4] S.H. Zhong, X. Huang and Z. Xiao, "Fine-art painting classification via two-channel dual path networks," *International Journal of Machine Learning and Cybernetics*, 2019, pp. 1-16.

[5] O. Kelek, N. Calik and T. Yildirim, "Painter Classification Over the Novel Art Painting Data Set via The Latest Deep Neural Networks," *Procedia Computer Science*, no. 154, 2019, pp. 369-376.

[6] A. Paul and C. Malathy, "An Innovative Approach for Automatic Genre-Based Fine Art Painting Classification," *In Advanced Computational and Communication Paradigms, Springer, Singapore*, 2018, pp. 19-27.

[7] E. Ohno, "State of the Fine art in the age of artificial intelligence," 2019, pp. 175-175.

[8] Y. Deng, F. Tang, W. Dong, F.Wu,O. Deussen, and C.Xu."Selective clustering for representative paintings selection, "*Multimedia Tools and Applications*, 2019, pp. 1-19.

[9] T.E. Lombardi, "The classification of style in fine-art painting," *Pace University*,2005.

[10] R.S. Arora, and A. Elgammal, "Towards automated classification of fine-art painting style: A comparative study," *In Proceedings of the 21st International Conference on Pattern Recognition*, 2012, pp. 3541-3544.

[11] G. Carneiro. N. P. da Silva, A. Del Bue and J. P. Costeira, "Artistic image classification: An analysis on the print art database," *In European Conference on Computer Vision, Springer, Berlin, Heidelberg*, 2012, pp. 143-157.

[12] B. Saleh, K. Abe and A. M. Elgammal, "Knowledge Discovery of Artistic Influences: A Metric Learning Approach," In ICCC, 2014, pp. 163-172.

[13] Y. Bar, N. Levy and L. Wolf, "Classification of artistic styles using binarized features derived from a deep neural network," *In European conference on computer vision, Springer, Cham*, 2014, pp. 71-84.

[14] H. Takeda, S. Farsiu, and P. Milanfar, "Robust kernel regression for restoration and reconstruction of images from sparse noisy data,*" In 2006 International Conference on Image Processing*, 2006, pp. 1257-1260.

[15] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann and H. Winnemoeller, "Recognizing image style," *arXiv preprint arXiv*:1311.3715, 2013.

[16] N. S. Murugan and G. U. Devi, "Detecting streaming of Twitter spam using hybrid method," *Wireless Personal Communications*, vol. 103, no. 2, 2018, pp. 1353-1374.

[17] N. S. Murugan, and G. U. Devi, "Feature extraction using LR-PCA hybridization on twitter data and classification accuracy using machine learning algorithms," *Cluster Computing*, 2018, pp. 1 3965-13974.

[18] N.S. Murugan, and G.U. Devi, "Detecting spams in social networks using ML algorithms-a review,*" International Journal of Environment and Waste Management*, vol.21, no.1, 2018, pp. 22-36.

[19] G. Manogaran, C.Thota,and D.Lopez, "Human-computer interaction with big data analytics," *In HCI challenges and privacy preservation in big data security*,2018, pp. 1-22, IGI Global.

[20] Y. Bar, N. Levy, and L. Wolf, "Classification of artistic styles using binarized features derived from a deep neural network," *In European conference on computer vision*, 2014, pp. 71-84, Springer, Cham.A.

[21] Oliva, and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope,*" International journal of computer vision*, vol.42, no. 3, 2001, pp. 145-175.

[22] L. Torresani, M. Szummer, and A.Fitzgibbon, "Efficient object category recognition using classemes," *In European conference on computer vision*, 2010, pp. 776-789, Springer, Berlin, Heidelberg.

[23] A. Vedaldi, and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," *In Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 689-692.

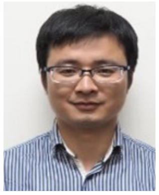
[24] A. Prathik, J. Anuradha, and K. Uma, “A Novel Algorithm for Soil Image Segmentation using Colour and Region Based System,” *International Journal of Innovative Technology and Exploring Engineering*, 2019, vol. 8, no. 10, pp.3544-3550.

[25] Z. Duric, W.D. Gray, R. Heishman, F. Li, A. Rosenfeld, M.J. Schoelles, C. Schunn, and H. Wechsler, “Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction,” *Proceedings of the IEEE*, 2002, vol. 90, no. 7, pp.1272-1289.

[26] C. Cedras and M. Shah, “Motion-based recognition a survey”, *Image and vision computing*, vol. 13, no. 2, 1995, pp. 129-155.

[27] R. Peterson, “Crossing Bridges That Connect the Arts, Cognitive Development, and the Brain,” *Journal for Learning through the Arts*, Vol. 1, no. 1, 2005, pp.2.

[28] E. Gredler Margaret and c.c. Shields. “Vygotsky’s legacy: A foundation for research and practice”. *Guilford Press*, 2008.

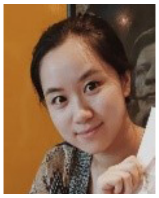


Fei Xu

Fei Xu, Ph.D., Assistant Researcher at Tsinghua University. His research interests are Cultural heritage preservation and exhibition design, relationship between graphic analysis and humanities. Email: xufeizhen@mail.tsinghua.edu.cn

Tong Wu

Tong Wu completed her Visual Communication at Beijing Institute of Fashion Technology, China. And she completed her Master Degree Design Theory at China Academy of Art, China. E-mail: 714029974@qq.com



Shali Huang

Shali Huang, Ph.D of Drama and Film. Postdoctoral of Art History. She working in the Tsinghua University and Central Academy of Drama, China. Stage designer for Drama, Interactive exhibition, Festival etc. Her researcher interests include modern art history, contemporary aesthetics, stage designers, vision of space and multimedia, etc. Curator for International Stage Art Network, such as International Stage Design Academic Week (2017-2019), World Stage Artist Palmstierna-Weiss Gunilla Exhibition (2017), International Stage Costume and Makeup Design Competition and Exhibition (2018), etc. Translated the book Art since 1900, World Scenography, 2013-2014 World Stage Art Student Exchange Works. Author of Encyclopedia of China.

Kuntong Han

Kuntong Han, Academy of Arts and Design, Tsinghua University, Beijing, China. Email: hankt16@qq.com

Wenwen Lin

Wenwen Lin, graduating in July with a degree of Master of Fine Arts, in Major in Fine Arts, Tsinghua University and Graduated in July with the degree of bachelor of fine arts, in Major in Fine Arts, Beijing Institute of Graphic Communication, Email: lww18@mails.tsinghua.edu.cn

Shizhong Wu

Shizhong Wu, Ph.D., Professor at Tsinghua University, Academic advisor of graduate students, Senior engineer. Director of the Art Institute of Urban Construction, Department of Arts and Design, Deputy Director of the Institute of Art Exhibition. His research interests are Digital art and design exhibitions, Cultural heritage preservation and innovation in technological integration, Digital art design. Email: wushiz@vip.sina.com

Sivaparthipan C.B

Dr C.B.Sivaparthipan is working as Assistant Professor in the Department of Computer Science & Engineering at Adhiyamaan College of Engineering, India. His area of interest includes Big Data Analytics, Internet of Things and Healthcare. For his credential he published more than 20 papers in refereed international journals and 13 papers in conferences. He also served as reviewer in many peer reviewed journals. Who also holds membership in Many Professional Bodies like ISTE, IAENG, IEEE etc., and also delivered many Guest lecturers in the Recent Field, conducted and participated in many social events like planting trees, educating school students, Clean India (Swachh Bharat) Scheme. He was also involved in helping Chennai flood rescue and got appreciated by YRC Coimbatore Division. Email: sivaparthipanece@gmail.com

Dinesh Jackson Samuel R

Dinesh Jackson Samuel is currently working as Associate Lecturer at Oxford Brookes University, UK. He also completed his Postdoctoral research from Oxford Brookes University, UK. Doctorial degree from Vellore Institute of Technology. Master of Engineering degree in Computer Science and Engineering from Anna University, Chennai in 2013. He has completed his Bachelor of Engineering in Computer Science and Engineering from Anna University, Chennai in 2010. Previously he worked as Assistant Professors in few renowned Institutions across India. He also worked as EFT POS terminal programmer in Marshal Equipment & Trading Co., Dubai. He worked on many research projects including identification of Tuberculosis bacilli from the microscopic sputum smear image/video, acquired using an automated microscopic stage. He has also completed certification course on Fundamentals of Deep Learning for Computer Vision issued by NVIDIA deep Learning Institute, Mainframe Application Programming issued by Maples, Certified D-Link trainer on SCT-Switching issued by D-Link. He has published his research work in various reputed impact factor journals and conference. His research interest areas include Image and Video Processing, Soft Computing, Deep Learning and Computer Vision. Email: rsamuel@brookes.ac.uk

Design of a Virtual Assistant to Improve Interaction Between the Audience and the Presenter

S. Cobos-Guzman[1]*, S. Nuere[2], L. De Miguel[1], C. König[3]

[1] Universidad Internacional de la Rioja, Escuela Superior de Ingeniería y Tecnología (ESIT), Logroño, La Rioja (Spain)
[2] Universidad Politécnica de Madrid, UPM, Escuela Técnica Superior de Ingeniería y Diseño Industrial. Madrid (Spain)
[3] Universitat Politècnica de Catalunya, UPC, BarcelonaTech. Barcelona (Spain)

Received 6 August 2020 | Accepted 27 June 2021 | Published 16 August 2021

Abstract

This article presents a novel design of a Virtual Assistant as part of a human-machine interaction system to improve communication between the presenter and the audience that can be used in education or general presentations for improving interaction during the presentations (e.g., auditoriums with 200 people). The main goal of the proposed model is the design of a framework of interaction to increase the level of attention of the public in key aspects of the presentation. In this manner, the collaboration between the presenter and Virtual Assistant could improve the level of learning among the public. The design of the Virtual Assistant relies on non-anthropomorphic forms with 'live' characteristics generating an intuitive and self-explainable interface. A set of intuitive and useful virtual interactions to support the presenter was designed. This design was validated from various types of the public with a psychological study based on a discrete emotions' questionnaire confirming the adequacy of the proposed solution. The human-machine interaction system supporting the Virtual Assistant should automatically recognize the attention level of the audience from audiovisual resources and synchronize the Virtual Assistant with the presentation. The system involves a complex artificial intelligence architecture embracing perception of high-level features from audio and video, knowledge representation, and reasoning for pervasive and affective computing and reinforcement learning to teach the intelligent agent to decide on the best strategy to increase the level of attention of the audience.

Keywords

AI, Human-machine Interaction, Learning, Multimedia, Virtual Assistance.

DOI: 10.9781/ijimai.2021.08.017

I. Introduction

In the academic world, the mission of the professor/teacher/lecturer (presenter) is to communicate knowledge in an efficient way [1]. This general objective can be achieved thanks to the experience of the presenter and the own skills acquired during the years of teaching. However, even if the presenter has great strategies for communication, it is difficult to get the attention of students or the public when interest lacks in the presentation's topic [2]. Feedback about the attitude and level of attention of the audience could be useful for the presenter to adapt and personalize the presentation according to the requirements of the audience.

Nowadays, there are several applications of virtual agents [3]-[5] or autonomous robots that can interact with the public using virtual or mechatronic faces [6]. Social interaction between humans and robots is paramount in applications where a high level of collaboration is required such as in hospitals or industry. For this reason, the design of a robot to generate an instantaneous positive reaction in humans is key for fostering human-robotic interaction.

Habitually, roboticists put much effort into the design of a new robot on physical embodiment properties such as locomotion, manipulation, haptic interactions, and face mechatronic interaction. However, for social interaction, technologies such as virtual agents or chatbots which focus on communication abilities rather than physical embodiments have been developed.

Therefore, this study presents a novel design of a virtual agent using artificial intelligence methods for creating a framework for effective collaboration between the public and the presenter. This framework is thought for improving the interaction of presentations in large venues (e.g., auditoriums with 200 people) where direct visual contact of the presenter and the audience is difficult. Thus, the new virtual agent can be used for: virtual interaction, preliminary design for a new future mechatronic system and to provide a hybrid combination of humans and virtual assistants that can help to increase the level of attention during a presentation.

The contribution of this research is the design of a Virtual Assistant with four levels of interaction to increase the audience's attention during a presentation. The four levels of interaction are derived from the intelligent system proposed. The intelligent system must detect different behavioral patterns of the audience such as divert/distract attention, people asleep, positive, or negative participation, and confused participation.

* Corresponding author.

E-mail address: salvador.cobos@unir.net

Furthermore, the study presents a set of psychological analyses of the virtual assistant's graphic design to validate the suitability for its use in the different levels of interaction. The psychological analysis considers the level of education and type of public to recommend the use of the virtual assistant with a given level of interaction. The information about the audience's profile can be used as an input in the system for adapting the mode of interactions.

This work also presents a design of the system's architecture, which is based on 3 different modules as follows: i) the first module is based on video and image recognition in order to identify different patterns (e.g., type of public and level of attention); ii) the second module analyzes audio and will contain natural language processing algorithms to recognize the phrases of the presenter and the feedback information from the public. iii) the third module contains an intelligent agent that can guarantee in real-time the synchronization with the presenter and the presentation; Moreover, the third module will depend on the recognition of patterns from the first and second module (perception modules) to correctly synchronize the collaboration between the presenter and the virtual assistant. For this, the system involves a knowledge representation and reasoning module to attribute a semantic meaning to the high-level perceptions regarding the attention and emotion. Knowledge representation is important to model the context of the current state of the world so that intelligent decisions can be taken. Reinforcement learning is proposed as the solution to teach the intelligent agent to decide on the best interaction with the audience at each moment.

For example, information such as applauds, and the expression of faces captured through image analysis provide feedback about the attitude of the audience in each moment, so that the Virtual Assistant can decide to take certain actions to improve the level of attention, i.e., activate a given interaction mode of the Virtual Assistant.

The remaining part of this article is organized as follows: Section II describes the design of the Virtual Assistant tackling the visual appearance followed by a discussion of the proposed design and the results of the evaluation of the virtual assistant animations by different publics (Section III). Section IV describes the design of the artificial intelligence system. Conclusions and future lines of work are presented in Section V.

II. Design of the Virtual Assistant

A. Design of the Appearance

During the last years, many virtual assistants have been designed with different shapes and voices, both in the academy and in the industry. Some virtual agents as Eliza [7] have feminine aspects, trying to evoke human characteristics. But trying to replicate human-like is not as easy and is not all about just benefits as there are so many consequences related to psychology, specifically with how we perceive and interact within the perception.

There are many studies regarding the concept of "uncanny valley". This concept explores the idea that there is something strange in the level of anthropomorphism of something that is not alive, in a biological way, but looks like. Professor Masahiro Mori [8] introduced this concept into robotics, after Sigmund Freud's article 'Das Unheimliche' [9]. The original term comes from Ernst Jentsch [10]. The creation of intelligent agents that work efficiently and effectively 'would be impossible to solve without understanding and using mechanisms of social-emotional cognition' [11]. It is important to keep in mind that the way how we perceive will incise directly in our interaction. In this context, we have tried to avoid this kind of perception though creating a virtual interaction.

A study compared three different animated objects during a cognitive task and their impact on the behavior and performance of primary school children [12]. The results revealed that animated objects were well-accepted as interacting partners and had a positive impact on their emotional state. Also, it shows that 3D objects (animal and robot) with more "live" characteristics elicited more positive behaviors, and the animal-like object decreases attention. Other studies from this field of research encourage managers to take care of the appearance and behavior of robots and promote collaboration between managers and researchers to define the limit of anthropomorphism [13]. Considering all this information, we have developed several designs for the robotic appearance and behavior of a Virtual Assistant aimed to improve the attention of the audience during a presentation.

B. Personification

In the XXI century, the resource of personification is one of the most used rhetorical figures in the field of animation and character design, but historically it has been used in the field of literary-fabulous creation and 2D animation (cartoons), since its inception. This resource has enormous potential to evoke empathy and closeness with the personified object, but at the same time, in the words of Radoslav (1996), the personification can "constitute an educational instrument, capable of producing profound positive effects" [14].

First, an analysis is made of the objects that may be suitable for the personification of the assistant. The training context in which the project is registered is online, therefore, suitable hardware elements are chosen for this purpose. Of all those available, the webcam is chosen for two compelling reasons. The first is the element that visually connects the participants in a virtual face-to-face educational context. Second, because it is considered that its nature is ideal to find formal and conceptual aspects that support the feeling of life that you want to give it.

C. Form

Trying to follow the advice of Andrew Jimenez, from Pixar Animation Studios [15], about the importance of not getting complicated when creating a character, since the important thing is that the idea in your head is raised simply and naturally [16], of all the types of webcams on the market, inspiration is chosen that model whose form is simple and at the same time versatile for its reorientation as "living being".

A model with a round central body is chosen, with the objective cantered, side light pilot on and base in one piece. From there, they begin to make sketches to choose which of their attributes are potentially suitable to support their mobility and characterization when presenting emotionally to the assistant.

At first, the objective is considered as a mouth adding the eyes and eyebrows. It is also necessary to add some element that supports the sensation of life and its expressive load, for which the power cable of the webcam is used as an arm that helps to emphasize the expressiveness of the assistant and his attitude in each of the poses. This aspect has special relevance as it already happened with other references from the world of animation, consulted for this design, such as, for example, the character in the movie Monsters, INC. [16], Mike Wazowski who was originally designed as a spherical body with two legs, without arms. But its creators realized that the arms would give him that feeling of reality and more suitable mobility to interact with his co-stars, resulting in closer and less strange to the viewer.

But the facial result is overloaded, losing the feeling of a webcam in which the most important thing is the objective. Thus, it is decided to simplify the model, turning the objective of the inspiration model into the eye of the assistant. Thus, the base of the webcam is modelled on the different attitudes that the assistant must simulate the movement of the feet.

D. Color

"No color is meaningless" [17], and for this reason, the colors that will be part of the assistant have been specifically chosen.

The wizard is white, with shades of blue. The lights and shadows in their movements will give rise to shades of gray. White color favors attention to the object, freeing itself from all colors. It is a bright color and favors the contrast in combination with blue color. Associated as Eva Heller [17] indicates to what is empty and light, it will highlight the expressions of the assistant.

For its part, the assistant's eye is blue on a white background, which according to Eva Heller is associated with intellectual qualities. Its typical combination is blue and white, both being related to intelligence, science, and concentration. Likewise, and according to the survey carried out by this author, the blue color is the most appreciated with 45%. Another fact to consider is the symbolism associated with the combination of different colors, and specifically blue, white, and gray, in this case, produced by the shadows produced by white, is associated with intelligence. Fig. 1 shows the preliminary sketches of the virtual assistant's concept.

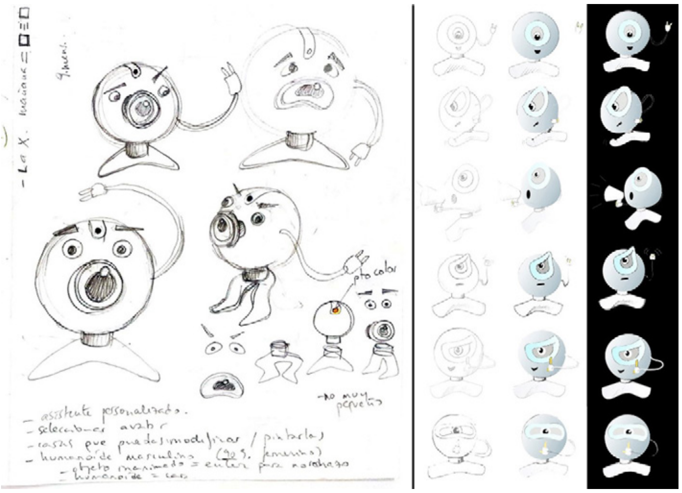

pto color
-no muy pequeño
- asistente personalizado.
- seleccionar avatar
- cosas que puedas modificar / pintarlo)
- humanoide masculino (90%. femenino)
- objeto animado = entre para
- humanoide =

Fig. 1. Preliminary sketches of the virtual assistant's conceptualization.

III. Design of the Virtual Behavior

According to [18], the emotional representation is composed of emotional characteristics, attributes, and attitudes, among others. In this sense, the assistant has an outgoing and close personality, capable of expressing negative or positive attitudes based on the different animations according to the objective to be achieved with each interaction with those attending virtual face-to-face sessions.

Regarding the level of attention in the audience, we consider four states: normal conditions; keep silence; low level of distraction, and high level of distraction. Based on this categorization, we try to reinforce the desired behaviors of the audience: High attention and interaction are rewarded with positive feedback and situations of distraction should be penalized through negative feedback, in this case, "anger face" (Fig. 2 (d)). Also, regarding the condition of interactions within the audiences, we present 4 states: people semi-sleeping; people sleeping, time for questions, and confusing questions. The four designs are clearly differentiating these states and represent a human-like reaction to each one (Fig. 3). About the condition of the responses of the presentation, we distinguish 2 states: positive feedback and positive participation. The virtual design reflects each state using common symbols (Fig. 4).

In the following, we explain how the virtual assistant controls the level of attention of the audience. The first mode is activated when the algorithm detects that people are talking or there are divert of attention. Fig. 2 (a) and (b) show the transition when the intelligent system detects when people are talking in the room, indicating that is important to keep silent in the room. This interaction is useful for the presenter to control the level of attention. Fig. 2 (c) and (d) show interactions when the algorithm detects different levels of distraction.

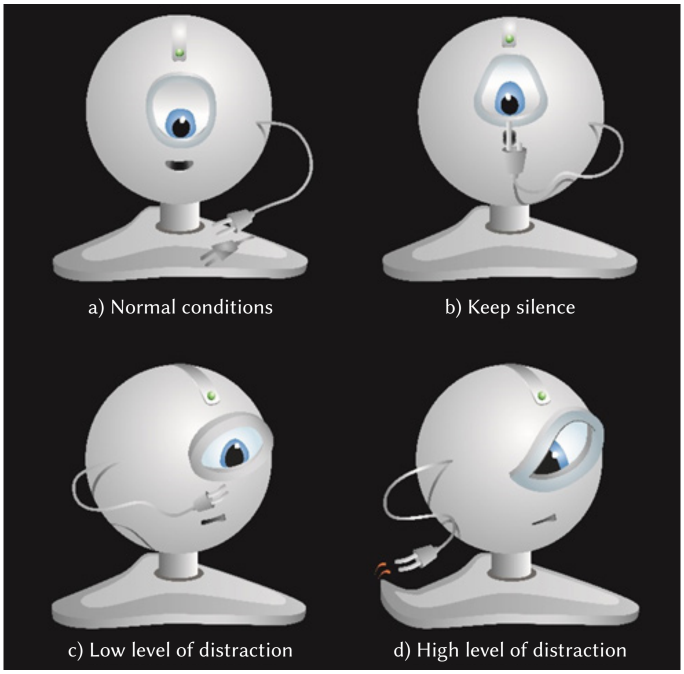


Fig. 2. Modes of interaction; (a) normal conditions; (b) keep silence; (c) low level of distraction and (d) high level of distraction.

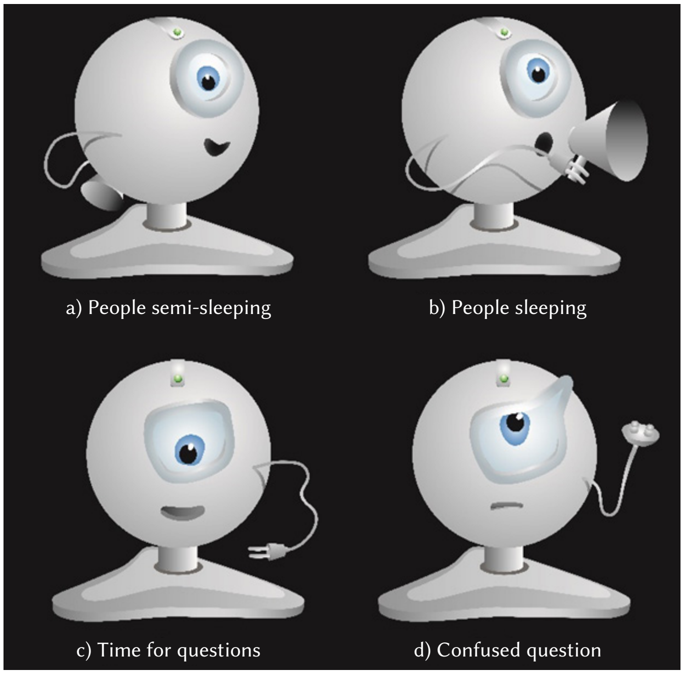


Fig. 3. Modes of interaction; (a) people semi-sleeping; (b) people sleeping (c) time for questions and (d) confused question.

The second mode is activated when the algorithm detects that people are sleeping in the room. Fig. 3 (a) and (b) show the transition when with the help of the presenter the level of attention can be increased.

Moreover, Fig. 3 (c) and (d) show the transitions when people are interacting or asking questions and the question is confusing. Finally, the last mode of interaction is when the system recognizes a positive interaction or comments as is shown in Fig. 4 (a) and (b).

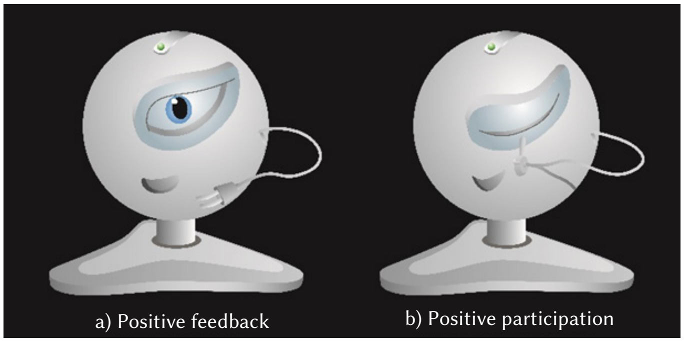


Fig. 4. Modes of interaction; (a) positive feedback; (b) positive participation

A. Discussion of the Virtual Design

Humans use different types of communication in verbal, written, and facial expressions. Face expressions [19] are enormously powerful because they instantaneously communicate emotions and intentions. This characteristic is the key aspect of the design of our artificial virtual assistant. For this reason, we use a smile and a gesture that represents a positive emotional state such as "it is O.K." represented in Fig. 4 referring to positive interactions or gestures like Fig. 2 (b) that transmit messages such as "keep silence".

Hence, the design presented in this work has been selected according to the research carried out in section II and III. This analysis helped to encounter the best features for a virtual assistant. Therefore, as a result, our proposed design is a non-anthropomorphic form, as we pretend to reduce the uncanny valley effect.

As a result, the Virtual Assistant's design incorporates a mixture of robotic and artificial life accessory characteristics. The design includes non-artificial characteristic or "live" characteristics such as the eye, mouth, and the cable-arm to give it an acquainted appearance. With these live characteristics, we try to increase the level of positive behavior in the public. This relies on the fact that humans can recognize patterns of facial expressions nearly instantly as the human learning system has been trained by years for face recognition. Therefore, this natural way of recognition helps to interpret a smile as a positive answer (Fig. 4) and a negative answer as the "anger face" (Fig. 2 (d)).

The color chosen for the principal object (the camera) is white as it represents purity, elegance, and truthfulness. It is also simple and recognizable. But also, it is the counterpart of the color black, and that is why there are some nuances of grey color to emphasize the object. The combination of these colors fosters the contrast of the image, highlighting the color chosen for the blue eye. As Molly E. Holzschlag [20] points out "as white is necessary for contrast and design, it is ideal to mix it with another color that has a stronger and more obvious meaning". The blue color inspires affection, friendship, confidence, and harmony. According to a survey [17], blue is the color that has more followers (46% men and 44 women) and only a few people do not like it (1% men and 2% women). It is commonly related to positive feelings.

The spherical form is directly in consonance with common traditional web cameras. It represents perfection as it does not have a beginning and an end, as well as it depicts protection and movement [21]. The shape of the figure also inspires stability as it incorporates a foot stand.

These positive and negative interactions are used during the presentation as input to a reinforcement learning algorithm designed to improve the quality of the presentation. Thus, the proposed architecture can activate the specific modes of interaction as a response to the detection of certain situations employing voice and vision recognition. Therefore, the combination of the different types of interaction with the presenter is a useful tool to support efficient communication of knowledge to the public.

Reinforcement learning has its foundation in psychological principles and practice. This methodology is based on positive reinforcement and punishment (showcased by the virtual assistant). For its implementation, it is important that the intelligent agent of the system can use functions to distinguish the audience's profile, which is derived from the recognized patterns in the audience through vision and sound processing technologies.

Additionally, the proposed solution represents a creative approach to define the boundaries between biological and artificial life from human expressions for a learning enhancement system avoiding the uncanny valley effect in the design of the virtual assistant.

As we explained in section II, animal representations are biological representations, but they generate negative effects because the audience may get more distracted. In contrast, if the design is too artificial it is possible that people cannot understand its meaning. For this reason, the presented design uses a mixture of features from biological and artificial representations to generate a better interaction with the audience. At the same time, this system will help the presenter to emphasize efficiently the knowledge during the presentation.

B. Evaluation of the Animations

The animations of the virtual design were evaluated using a psychological study based on a 'discrete emotions questionnaire' [22]. In this analysis, 34 participants (23 female and 11 male) with an age between 20 and 50 participated in the evaluation of the five animations as follows:

Animation 1: transition from Fig. 2 (c) and 2 (d); animation 2: transition from Fig. 2 (a) and 2 (b); animation 3: transition from Fig. 3 (a) and 3 (b); animation 4: transition from Fig. 4 (a) and 4 (b) and animation 5: transition from Fig. 3(c) and 3(d).

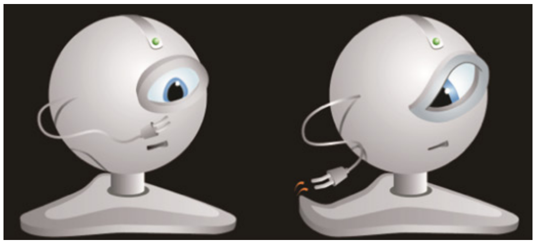
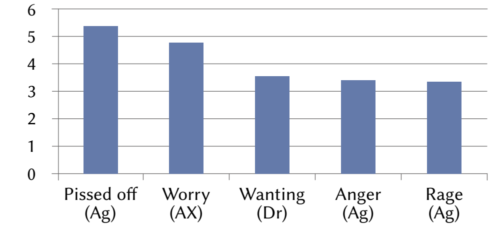
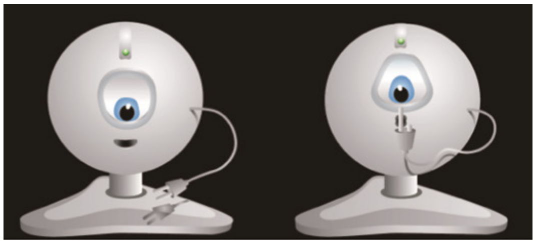
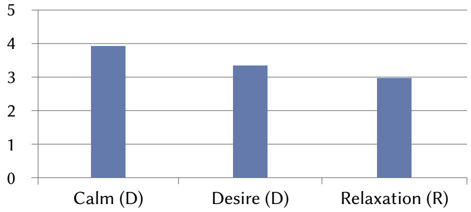
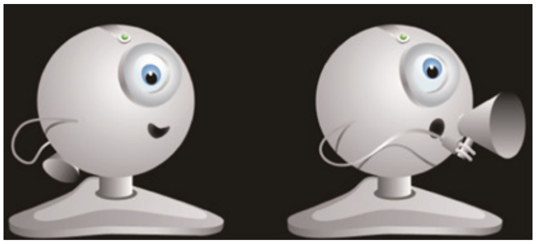
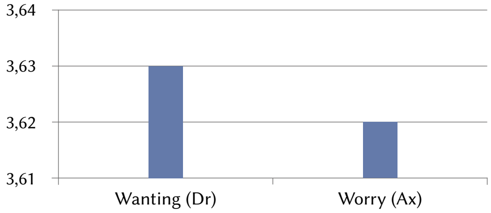
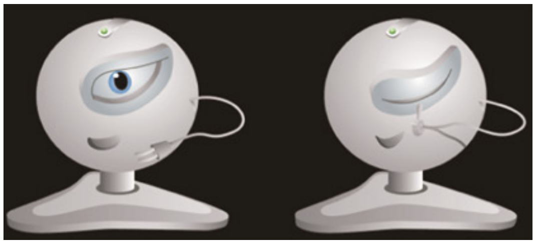
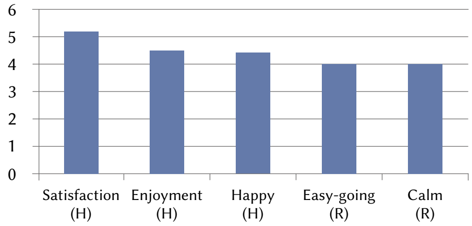
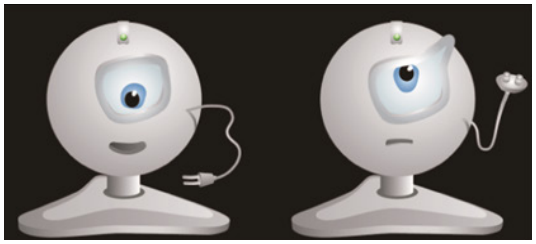
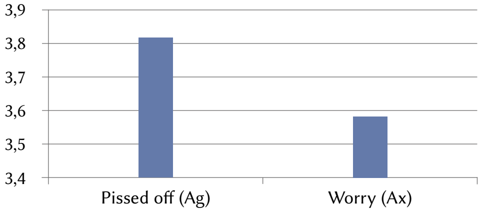
The questionnaire is based on different items that can measure the following emotions: Anger (Ag), Anxiety (Ax), Desire (Dr), Disgust (Dg), Fear (F), Happiness (H), Relaxation (R) and Sadness (S). The emotions were measured on a scale ranging from 1 to 7. The results of the questionnaire are shown in Table I and explained below according to the most relative scores (from 3).

The results of animation 1 generate three emotions of anger as highest values, 1 emotion of anxiety, and 1 emotion of desire. Therefore, this animation achieves the main objective of generating a negative emotion in the public. Animation 2 produced two emotions of relaxation and one of desire for the keep silent animation (Fig. 2 b).

Animation 3 resulted in emotions of desire and one of anxiety for the case 'people are sleeping' (Fig. 3 b). Animation 4 yielded three emotions of happiness and two of relaxation. This result is important because the positive feedback animation is evaluated by the public through a combination of positive emotions such as happiness and relaxation. Finally, the animation 5 generates two emotions of Anger and Anxiety.

As conclusion, the results of the survey confirm that the emotional reactions of the people coincide with the emotional effect, which was designed for each animation. These results are important because if any animation does not generate an adequate emotional reaction in the audience, the interaction with the public can be ineffective, resulting in frustration and ignoring of the virtual assistant.

TABLE I. Evaluation of the Five Animations

Animation 1		
Animation 2		
Animation 3		
Animation 4		
Animation 5		

IV. Design of the Artificial Intelligence Architecture

The artificial intelligence architecture for the human-machine interaction system comprises three main modules. Two perception modules are responsible for the visual recognition and acoustic analysis (sound analysis and natural language recognition) of patterns from the audience and a third module, responsible for the real-time synchronization between the presenter and the virtual assistant. Fig. 5 describes the organization of the main modules.

A. Perception Modules

The purpose of Module 1 is the recognition of the public's attitude towards the presentation through the caption of the visual focus of attention. A similar approach has been used recently to analyze the behavior of persons in group meetings [23] using a multi-modal sensor approach. In our approach, the visual patterns are used to detect situations remotely where the public is not putting the appropriate attention. This can be implemented by using face recognition algorithms, which detect whether the eyes are open or closed [24] or calculate the eye gaze direction [25]. These two variables allow the system to derive information about the level of attention of the audience. The eye gaze direction and head movements have resulted in the key for detecting attention shifts in a previous educational study [26]. Analyzing the eye gaze direction, the system can determine automatically how many people pay attention to the projection of the presentation. Moreover, module 1 must recognize the type of people according to the average age of the audience [27]. Thus, these types of techniques allow the system to determine if people are asleep, divert, or distract attention.

On the detection of these previous situations, the virtual assistant will be activated with the corresponding interaction model, as described in section III.

Module 2 is used to capture the acoustic feedback from the audience. From a technical point of view, the system must carry out acoustic analysis to recognize certain human activities related to different types of noise. Acoustic scene and event recognition are an

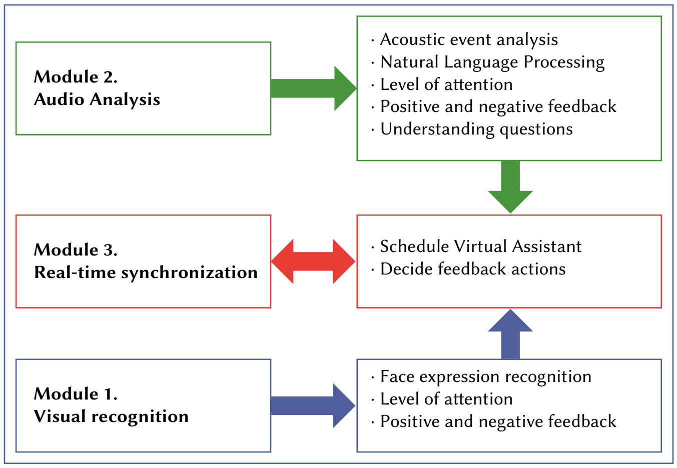

Module 2.
Audio Analysis
· Acoustic event analysis
· Natural Language Processing
· Level of attention
· Positive and negative feedback
· Understanding questions
Module 3.
Real-time synchronization
· Schedule Virtual Assistant
· Decide feedback actions
Module 1.
Visual recognition
· Face expression recognition
· Level of attention
· Positive and negative feedback

Fig. 5. System architecture of the intelligent system for the virtual assistant.

active field of research, where deep neural network models currently outperform humans in recognizing certain types of acoustic events [28]. In the present system, the focus is on the detection of attention drift primarily based on the level of noise in the room [29]. Another feature of the acoustic module is speech recognition to analyze comments from the audience. Speech recognition relies on complex models for natural language processing [30-32]. Active research during the last decades, especially in the area of neural networks [33] have implemented speech recognition as a helpful human-machine communication technology [34] in virtual assistants. In our system the automatic recognition of speech is designed as an input to the content of the presentation as the algorithm can detect automatically in which slide of the presentation the comment is made. Another functionality of the module is the detection of positive and negative comments and other types of feedback (such as applause). This information will be used for the intelligent agent of module 3 to decide the best mode of the virtual assistant.

The integration of information from several signals (video, audio, or linguistic) is referred to as multimodal feature extraction and fusion providing more reliable and rich information compared to that derived from single-modality signals [35]. Nevertheless, the high dimensionality and the complexity of the input pattern require sophisticated machine learning models for the extraction of high-level features from raw data. Recently, several cloud providers specialized in artificial intelligence technologies, such as Google or IBM, offer services for the automatic extraction of features from video, audio, and natural speech. These services are provided with a scalable and quite affordable cost, so that system developers may integrate such technologies in their software. Other alternatives are open-source software solutions, such as 'Pliers' [36], which provide an extensible framework to automatically implement semantic feature extraction from audiovisual resources.

B. Knowledge Representation

To process the perceptions from the visual recognition or acoustic module, the recognized perceptions must be transformed into treatable information for an intelligent agent by a proper knowledge representation. Semantic networks (or ontologies) provide a formal description of domain knowledge defining the characteristics of entities and the relationship between them. Ontologies use description logics such as OWL as the representational formalism for the domain description providing the ability to apply logical inference on the entities and model a context (state of the world). The ability of reasoning on knowledge representation is an enabler for pervasive (context-aware) computing [37].

Many knowledge bases have been released with the aim of information reuse and sharing. ConceptNet [38], for example, is an important knowledge base describing the semantic of several thousand of concepts. SenticNet [39] is an extension from ConceptNet with the ability to describe sentiment-based annotations. Such representation of human emotions in a computer interpretable format is tackled as affective computing [40]-[41], an interdisciplinary discipline, which aims to enable computer systems to recognize and interpret human emotions. SenticNet considers the categories of admiration, anger, disgust, fear, interest, joy, sadness, and surprise. Recently, OntoSenticNet [42] was released providing a connotative description of emotions to the concepts using polarity values. For the present system, we consider the use of OntoSenticNet, as emotions and attention levels must be attributed to the concepts sensed from the perception module.

The following examples illustrate the conceptualization of three relevant perceptions in the proposed system, namely 'doze_off' (tiredness), 'applause', and 'noise' by OntoSenticNet (Table II). The term 'doze_off' is described as semantically close to the terms 'fall_asleep', 'bed_time', 'become_bored' or 'drowsiness'. The corresponding mood tags are sadness and disgust, and the overall emotive valuation is negative with an intensity of 0.54. The emotional assessment (sentics values) gives a fine-grained evaluation of emotions in different categories. All three analyzed terms show a proper conceptualization regarding the attention and emotional categorization from a human point of view, which confirms the adequacy to use the OntoSenticNet as a knowledge base for emotion characterization in the system.

TABLE II. Conceptual Map of the Term 'Doze-off' (1), 'Applause' (2) and 'Noisy' (3)

N	Semantic	Mood tags	Sentics	Polarity
1	fall_asleep bed_time become_bored drowsiness	Sadness Disgust	Pleasant (0.48) Attention (-0.08) Sensitive (0.04) Aptitude (0)	Value: Negative Intensity: -0.54
2	watch_play entertainment cheering attend_	Interest Admiration	Pleasant (0.164) Attention (0.24) Sensitive (-0.14) Aptitude (0.28)	Value: Positive Intensity: 0.179
3	Loud uncontrollable obsolete last_clue	Sadness Disgust	Pleasant (-0.44) Attention (-0.15) Sensitive (0) Aptitude (-0.36)	Value: Negative Intensity: -0.21

C. Reasoning Module

Module 3 implements the artificial intelligence agent, who decides on the operating mode of the virtual assistant in real-time. According to [42] "all reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments". In this application, the agent uses the information from the visual and acoustic module as sensing information about the state of the audience. The actions on the audience are taken by deciding a given virtual assistance mode and the goal is to optimize the audience's attention. Thus, when positive answers are detected, the system will activate the positive feedback animation (Fig. 4). However, if a negative answer is detected, the system will activate the animation of "anger face" (Fig. 2 (d)). Moreover, the algorithm will decide which animation activate from negative to positive parameters. Fig. 6 shows a description of the information flow in this human-machine interaction system.

The agent uses a reasoning module to decide on the actions to take in the current state. Reasoning can be implemented with traditional logic-based symbolism using knowledge representation languages (OWL) or alternatively using machine-learning approaches. Symbolic reasoning applies logic inference to a set of rules and facts (instances in the ontology) deriving higher-level knowledge from the observations, which represent the context of the state. Recent research [43] focuses also on the use of neural networks, capable to project in embeddings the equivalent to the logical reasoning in ontologies. These approaches are interesting, as they overcome some commonplace shortcomings of logic-based reasoning in ontologies, such as sensitivity to noise and missing values.

The ontology to be used in this system must be carefully designed to cover generic descriptions from the behavior of the audience, attitude, level of attention, and questions of the audience about the presentation to model the context. The knowledge regarding emotions of OntoSenticNet is complementary to these representations of entities of other knowledge bases and can be integrated through different APIs in external systems.

Furthermore, the agent should be able to make intelligent decisions. For this, it is necessary to implement a reinforcement learning (RL) system, where the agent learns from its experience of interaction with the environment. In this application, a positive reward is given if the attention of the public increases, while a negative reward is given when the attention decreases.

In RL, an agent observes a state and takes actions, which generates a transaction of the current context to a new state, providing a reward to the agent as feedback of the transaction. The objective of the agent is to learn a strategy that maximizes the reward.

An overview of the literature shows that deep learning models are reported as the state-of-the-art technology for ‘enabling reinforcement learning to scale to problems that were previously intractable’ [44] as these models are capable to process high dimensional input patterns and complex state scenarios. Another interesting work is presented in [45], which gives an overview of recent advances in the integration of natural processing language models in reinforcement applications.

D. Discussion of the Artificial Intelligence System

The modules constituting the artificial intelligence architecture for the proposed human-machine interaction system were described as part of preliminary system design. The implementation of the system requires dedicated work to solve the respective tasks and it is part of the future lines of work.

V. Conclusions

This work has presented a novel graphic design of a Virtual Assistant with four levels of interaction. An artificial intelligence system has been designed to activate different interaction modes of a Virtual Assistant to improve the communication between the public and the presenter. The proposed system is designed for four levels of attention, such as normal conditions; keep silent; low level of distraction, and a high level of distraction. When one of these levels of attention is recognized by the intelligent agent, a positive or negative interaction of the virtual assistant is induced in order to increase the level of attention of the audience.

Moreover, the presented designs of the assistant rely on non-anthropomorphic forms with “live” characteristics (eye, mouth, and cable-arm). These features help the audience to automatically recognize situations without the need for an explicit explanation of the presenter (e.g., ‘keep silence’). This characteristic makes the system autonomous as the meaning of the interactions is intuitive and easy to understand for humans. An exception is the type of interaction when people are sleeping. In this case, the interaction of the presenter is required.

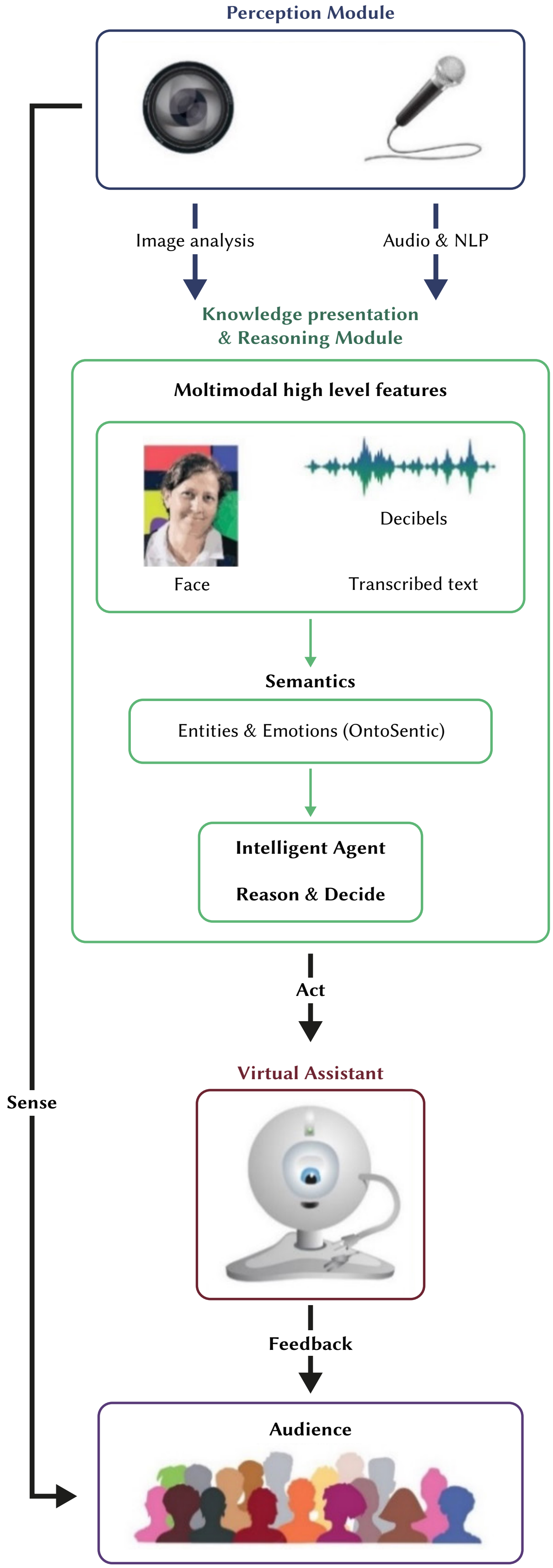


Fig. 6. Description of the human-machine interaction of the system.

Future lines of work will focus on the measurement of the impact of the proposed system on the quality of presentations. The objective is to measure the level of attention of the public comparing conventional ways of presentations and the incorporation of the virtual assistant.

A preliminary study of the intelligent architecture necessary to implement the Virtual Assistant in a human-machine interaction system has shown the need to use several modules. Perceptions of the environment are captured from an audio and visual recognition module extracting high-level features representing behavior or attitudes.

Ontology-based knowledge representation is proposed as a framework for the intelligent agent responsible for the activation of the virtual assistant. Reinforcement learning is recommended for deciding on the best strategy of the virtual assistant to achieve a high level of attention in the audience. The implementation of a proof of concept of the described framework is considered as a future line of work.

Acknowledgment

We would like to acknowledge Gema Fernández-Blanco Martín for her contribution on the psychological analysis of the survey results.

References

[1] B. Alters and C. Nelson, “Perspective: teaching evolution in higher education”, Evolution, vol. 56, no. 10, p. 1891, 2002, doi: 10.1554/0014-3820(2002)056[1891:pteihe]2.0.co;2.

[2] P.R. Pintrich, A. Zusho “Student Motivation and Self-Regulated Learning in the College Classroom,” in: *Smart J.C., Tierney W.G. (eds) Higher Education: Handbook of Theory and Research. Higher Education: Handbook of Theory and Research, vol 17.* Springer, Dordrecht, 2002, pp. 55-128, doi:10.1007/978-94-010-0245-5_2.

[3] A. Nijholt, “Towards the Automatic Generation of Virtual Presenter Agents”, *Informing Science: The International Journal of an Emerging Transdiscipline*, vol. 9, pp. 097-110, 2006, doi: 10.28945/474.

[4] R. Looije, M. Neerincx and F. Cnossen, “Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors”, *International Journal of Human-Computer Studies*, vol. 68, no. 6, pp. 386-397, 2010, doi: 10.1016/j.ijhcs.2009.08.007.

[5] C. Bartneck, J. Reichenbach, and A. van Breemen, “In Your Face, Robot! The Influence of a Character’s Embodiment on How Users Perceive Its Emotional Expressions”, in *Proceedings of the Design and Emotion*, Ankara, Turkey, 2004, pp. 32–51.

[6] W. Burgard et al., “Experiences with an interactive museum tour-guide robot”, *Artificial Intelligence*, vol. 114, no. 1-2, pp. 3-55, 1999, doi: 10.1016/s0004-3702(99)00070-3.

[7] DeixiLabs, Accessed: Feb. 12, 2020. [Online]. Available: http://www.deixilabs.com/eliza.html

[8] M. Mori, K. MacDorman and N. Kageki, “The Uncanny Valley [From the Field]”, *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98-100, 2012, doi: 10.1109/mra.2012.2192811.

[9] S. Freud, The Uncanny (1919). Accessed: Jun. 28, 2021. [Online]. Available: https://web.mit.edu/allanmc/www/freud1.pdf

[10] E. Jentsch, “On the psychology of the uncanny (1906)”, *Angelaki*, vol. 2, no. 1, pp. 7-16, 1997, doi: 10.1080/09697259708571910.

[11] A. Chubarov and D. Azarnov, “Modeling Behavior of Virtual Actors: A Limited Turing Test for Social-Emotional Intelligence”, in: *Samsonovich A., Klimov V. (eds) Biologically Inspired Cognitive Architectures (BICA) for Young Scientists. BICA 2017. Advances in Intelligent Systems and Computing, vol 636.* Springer, Cham, 2018, doi:10.1007/978-3-319-63940-6_5.

[12] V. André et al., “Ethorobotics applied to human behaviour: can animated objects influence children’s behaviour in cognitive tasks?”, *Animal Behaviour*, vol. 96, pp. 69-77, 2014, doi: 10.1016/j.anbehav.2014.07.020.

[13] S. Kim, B. Schmitt and N. Thalmann, “Eliza in the uncanny valley: anthropomorphizing consumer robots increases their perceived warmth but decreases liking”, *Marketing Letters*, vol. 30, no. 1, pp. 1-12, 2019, doi: 10.1007/s11002-019-09485-9.

[14] K. Sullivan, G. Schumer and K. Alexander, “Ideas for the animated short: finding and building stories” Focal Press, USA, 2008, pp. 64-67.

[15] K. I. Radoslav, “Televisión, dibujos animados y literatura para niños”, *Aisthesis*, 29, pp 33-49, 1996.

[16] Pixar Animation Studios, Accessed: Mar. 19, 2021. [Online]. Available: https://www.pixar.com/feature-films/monsters-inc

[17] E. Heller, “Psicología del color, Cómo actúan los colores sobre los sentimientos y la razón”, Gustavo Gili, Barcelona, 2004.

[18] J. Guzmán Ramírez, “Una metodología para la creación de personajes desde el diseño de concepto”, *Iconofacto*, vol. 12, no. 18, pp. 96-117, 2016, doi: 10.18566/v12n18.a06.

[19] I. Revina and W. Emmanuel, “A Survey on Human Face Expression Recognition Techniques”, *Journal of King Saud University - Computer and Information Sciences*, 2018, doi: 10.1016/j.jksuci.2018.09.002.

[20] M. E. Holzschlag, “Color para sitios web”, McGraw Hill, México, 2002.

[21] A. Frutiger, “Signos, símbolos, marcas, señales”, Gustavo Gili, México, 2007.

[22] C. Harmon-Jones, B. Bastian and E. Harmon-Jones, “The Discrete Emotions Questionnaire: A New Tool for Measuring State Self-Reported Emotions”, *PLOS ONE*, vol. 11, no. 8, p. e0159915, 2016, doi: 10.1371/journal.pone.0159915.

[23] I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, and R. Radke, (2018). “A multimodal-sensor-enabled room for unobtrusive group meeting analysis, in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 347-355.

[24] G. Trotta et al., “A neural network-based software to recognise blepharospasm symptoms and to measure eye closure time”, *Computers in Biology and Medicine*, vol. 112, p. 103376, 2019, doi: 10.1016/j.compbiomed.2019.103376.

[25] Y. Wang, R. Huang and L. Guo, “Eye gaze pattern analysis for fatigue detection based on GP-BCNN with ESM”, *Pattern Recognition Letters*, vol. 123, pp. 61-74, 2019, doi: 10.1016/j.patrec.2019.03.013.

[26] Y. Kuo, J. Lee and M. Hsieh, “Video-Based Eye Tracking to Detect the Attention Shift”, *International Journal of Distance Education Technologies*, vol. 12, no. 4, pp. 66-81, 2014, doi: 10.4018/ijdet.2014100105.

[27] M. Shakeel and K. Lam, “Deep-feature encoding-based discriminative model for age-invariant face recognition”, *Pattern Recognition*, vol. 93, pp. 442-457, 2019, doi: 10.1016/j.patcog.2019.04.028.

[28] J. Abeßer, “A Review of Deep Learning Based Methods for Acoustic Scene Classification”, *Applied Sciences*, vol. 10, no. 6, p. 2020, 2020, doi: 10.3390/app10062020.

[29] Y. Li, Q. He, S. Kwong, T. Li and J. Yang, “Characteristics-based effective applause detection for meeting speech”, *Signal Processing*, vol. 89, no. 8, pp. 1625-1633, 2009, doi: 10.1016/j.sigpro.2009.03.001.

[30] Y. Belinkov and J. Glass, “Analysis Methods in Neural Language Processing: A Survey”, *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 49-72, 2019, doi: 10.1162/tacl_a_00254.

[31] N. Saleem, M.I. Khattak, “Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 84-90, 2020, doi: 10.9781/ijimai.2019.06.001.

[32] N. Saleem, M.I. Khattak, E. Verdú, “On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 78-89,2020. doi: 10.9781/ijimai.2019.12.001.

[33] A. Torfi, R. Shirvani, Y. Keneshloo, N. Tavvaf, and E. Fox, (2020). “Natural Language Processing Advancements By Deep Learning: A Survey”. *ArXiv*, Vol. abs/2003.01200, n. pag., 2020, available at: https://www.arxiv-vanity.com/papers/2003.01200/

[34] A. Guzman, “Voices in and of the machine: Source orientation toward mobile virtual assistants”, *Computers in Human Behavior*, vol. 90, pp. 343-350, 2019, doi: 10.1016/j.chb.2018.08.009.

[35] M. Gurban, “Multimodal feature extraction and fusion for audio-visual speech recognition”. Lausanne, EPFL, 2009, doi: 10.5075/epfl-thesis-4292.

[36] Q. McNamara, A. De La Vega and T. Yarkoni, “Developing a comprehensive framework for multimodal feature extraction”, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, 2017, pp. 1567-1574.

[37] X. H. Wang, D. Q.Zhang, T. Gu, and H. K. Pung, (2004, March). “Ontology

based context modeling and reasoning using OWL", in *IEEE annual conference on pervasive computing and communications workshops*, Orlando, FL, USA, 2004, pp. 18-22.

[38] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge", in *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 2017, pp. 4444-4451.

[39] E. Cambria, "Affective Computing and Sentiment Analysis", *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102-107, 2016, doi: 10.1109/mis.2016.31.

[40] R. Picard, "Affective computing: challenges", *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55-64, 2003. doi: 10.1016/s1071-5819(03)00052-1.

[41] M. Dragoni, S. Poria and E. Cambria, "OntoSenticNet: A Commonsense Ontology for Sentiment Analysis", *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 77-85, 2018. doi: 10.1109/mis.2018.033001419.

[42] R. Sutton, F. Bach and A. Barto, "Reinforcement Learning", Massachusetts: MIT Press Ltd, 2018.

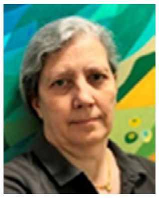
[43] P. Hohenecker and T. Lukasiewicz, "Ontology Reasoning with Deep Neural Networks", *Journal of Artificial Intelligence Research*, vol. 68, 2020, doi: 10.1613/jair.1.11661.

[44] K. Arulkumaran, M. Deisenroth, M. Brundage and A. Bharath, "Deep Reinforcement Learning: A Brief Survey", *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26-38, 2017. doi: 10.1109/msp.2017.2743240.

[45] J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel, "A Survey of Reinforcement Learning Informed by Natural Language", *ArXiv*, Vol. abs/1906.03926, 2019, available at: http://arxiv.org/abs/1906.03926

Salvador Cobos Guzman

He is professor of Robotics, Cyberphysical Systems, and Artificial Intelligence in the Faculty of Engineering at the Universidad Internacional de La Rioja, Spain. He teaches in the master's degree of Industry 4.0. He received his first B.Sc. in Industrial Robotics from National Polytechnic Institute (I.P.N.), Mexico, in July 2003. The second Engineering degree received was in Automation and Industrial Electronics (A.I.E.) from Polytechnic University of Madrid (U.P.M), Spain, in 2009. Also, he received an Advanced Studies Diploma (A.S.D.) in Robotics and Automation from Polytechnic University of Madrid (U.P.M), Spain, in 2007, and his Ph.D. with honors ("Sobresaliente Cum Laude") in Robotics and Automation from Polytechnic University of Madrid (U.P.M), Spain, in 2010. The Ph.D. thesis was focused on obtaining virtual human hand models for virtual manipulation tasks. Dr. Cobos worked as a postdoc in the UMI 375-LAFMIA laboratory for one year from 2011 to 2012. During this period, he was working in the design of underwater robots. Also, he was a Senior Research Fellow at The University of Nottingham working in the area of hyper-redundant robots for in-situ inspection from 2012 to 2016.



Silvia Nuere

She is a professor in the Department of Mechanical, Chemical and Industrial Design in the Technical School of Engineering in Industrial Design at the Universidad Politécnica de Madrid, Spain. She teaches Artistic Drawing, Basic Design, Graphic Design and Visual Communication. She received her bachelor's in fine arts in 1989 and her PhD in 2002 from the Universidad Complutense de Madrid, Spain. Her research interests include, teaching and learning methods based on Project Oriented Learning and in the need of mixing different fields of knowledge as art, design and engineering. She has promoted this approach to education through more than 30 Innovation Education Projects and from 2011 she is the Creator and Director of the scientific journal ArDIn (Art, Design and Engineering). She is author and co-author of more than 50 publications about artistic learning methods and humanistic approach to education. She has also, as an artist, took part in more than 20 collective exhibitions and made several illustrations for the Scientific Magazine "Investigación y Ciencia", the Spanish edition of the Scientific American Magazine.

Laura de Miguel

She has a Bachelor of Fine Arts from UCM in Image Arts and PhD in Fine Arts from the Universidad Complutense de Madrid, Spain. She is a professor in the Higher School of Engineering and Technology, at the Universidad Internacional de la Rioja (UNIR), Spain. With more than 15 years in the university field, she has specialized in being a teacher and author of content in areas of Graphic-artistic Expression and Design (graphic, industrial, fashion). She teaches subjects such as: creativity, drawing, analysis of the form or projects. She has numerous publications and has participated in artistic and academic outreach activities. Furthermore, in parallel to these activities, she has always kept her facet as a creator through the generation of multidisciplinary workshops (painting, drawing, engraving or short film.) shown in individual and group exhibitions. She has also been curator of exhibition projects, designed and directed art workshops in spaces for cultural dissemination, the web or congresses. Her lines of research focus on creative processes in Art-Design, relationship between Art-Design-Society, educational innovation in creative training, methodological exploration for the teaching of graphic representation, evaluation systems in areas of graphic expression, direction of audiovisual productions as an awareness tool.

Caroline König

She graduated in 2007 in Informatic Engineering at the Faculty of Informatics of the Polytechnic University of Catalonia and worked as software engineer during several years. In 2012 she received a master's degree in advanced Methods in Artificial Intelligence from the Universidad Nacional a Distancia (UNED). From 2013 - 2018 she was a predoctoral researcher of the PhD program of Artificial Intelligence of the UPC of the Soft Computing (SOCO) research group. In 2018 she received her PhD in Artificial Intelligence from the UPC. Since 2019 she is teacher of the Computer Science Department at Universitat Politècnica de Catalunya, UPC and postdoctoral researcher of the 'ML-PROMOLDYN' project (2020). She is involved in the development of artificial intelligence applications in the field of industry, bioinformatic and robotics. Her research interests are on deep learning approaches for automatic feature extraction from multimodal sequential data, anomaly detection and explainability of machine learning models.

Learning Analytics to Detect Evidence of Fraudulent Behaviour in Online Examinations

Antonio Balderas[1]*, Manuel Palomo-Duarte[1], Juan Antonio Caballero-Hernández[2], Mercedes Rodriguez-Garcia[3], Juan Manuel Dodero[1]

[1] Departamento de Ingeniería Informática, Universidad de Cádiz, Escuela Superior de Ingeniería, Puerto Real (Spain)
[2] EVALfor Research Group, Universidad de Cádiz, Puerto Real (Spain)
[3] Departamento Ingeniería en Automática, Electrónica, Arquitectura y Redes de Computadores, Universidad de Cádiz, Escuela Superior de Ingeniería, Puerto Real (Spain)

Received 26 February 2021 | Accepted 2 September 2021 | Published 27 October 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Lecturers are often reluctant to set examinations online because of the potential problems of fraudulent behaviour from their students. This concern has increased during the coronavirus pandemic because courses that were previously designed to be taken face-to-face have to be conducted online. The courses have had to be redesigned, including seminars, laboratory sessions and evaluation activities. This has brought lecturers and students into conflict because, according to the students, the activities and examinations that have been redesigned to avoid cheating are also harder. The lecturers' concern is that students can collaborate in taking examinations that must be taken individually without the lecturers being able to do anything to prevent it, i.e. fraudulent collaboration. This research proposes a process model to obtain evidence of students who attempt to fraudulently collaborate, based on the information in the learning environment logs. It is automated in a software tool that checks how the students took the examinations and the grades that they obtained. It is applied in a case study with more than 100 undergraduate students. The results are positive and its use allowed lecturers to detect evidence of fraudulent collaboration by several clusters of students from their submission timestamps and the grades obtained.

Keywords

Cheating, Evaluation, Learning Analytics, Learning Management System, Learning Records.

DOI: 10.9781/ijimai.2021.10.007

I. Introduction

The learning process in higher education is no longer conceivable without information technology. In particular, Learning Management Systems (LMSs) are a meeting point for students and faculties in the university, where faculties organise their courses and set up learning activities, while students find learning material and communicate with their lecturers. However, in terms of student assessment, lecturers usually prefer face-to-face rather than online examinations. The main reason for this is the concern that students can easily cheat in online examinations because the lecturers lose control of what the students do while taking their examinations [1].

This concern has become particularly significant in the context of the Covid-19 pandemic. Due to forced confinement in spring 2020, all Spanish students stopped attending classes and lectures, and moved from classrooms to video conferencing [2]. The same applied to evaluation activities, which were moved to the LMS through individual assignments, questionnaires and synchronous oral interviews, among others [3]. Oral examinations can be a solution to ensure the absence of fraud in a student's examination. Unfortunately, oral examinations are not always possible, either because they are not sustainable when there are too many students or because the course matter is not suitable for oral communication. On the other hand, examinations based on multiple-choice questionnaires were widely used as they are immediate to grade through their automated settings. However, they are prone to cheating by students [4].

To alleviate their concerns about student cheating during the confinement, lecturers could propose different examinations that were harder than those they usually set in face-to-face sessions (e.g., more difficult multiple-choice questions, shorter time to answer, etc.). In addition, the students sometimes reported that the LMS often suffered from connectivity problems because of the number of simultaneous connections taking place during the examination. These problems led to interruptions in the LMS service during the examinations, which would be a serious problem for the students involved. Because this issue was out of the lecturer's control, the students asked for the examinations to be asynchronous to alleviate this problem. Asynchronous examinations allow students to take the examination at different times [5].

This research focuses on the detection of cheating behaviour of students when asynchronously taking exams or submitting assignments. To get an insight into students who cheat in examinations, lecturers can check the student records on the LMS. Unfortunately, the information provided for large groups of students

* Corresponding author.
E-mail address: antonio.balderas@uca.es

is hard to manually analyse. Learning analytics support lecturers in both improving the assessment of their students and monitoring their learning process [6], [7].

The research question that arises from this context is: Can lecturers collect evidence of cheating students through LMS activity records? This study applies learning analytics techniques to help detect evidence of students who fraudulently collaborate in online examinations. We propose a process model to obtain evidence of how the students take examinations based on their submission timestamps and the grades obtained. For this purpose, we developed Py-Cheat, which is a software tool to identify evidence of fraudulent collaboration among students when performing online activities. It is applied in a case study with more than 100 undergraduate students who submitted several programming assignments throughout the course and later took an examination based on a multiple-choice questionnaire.

The rest of the paper is structured as follows. In the second section, we describe the background of this work. In the third section, we present the materials and methods used, including the Py-Cheat tool. The fourth section presents the results. The fifth section discusses the implications of this study. Finally, the conclusions of this study are drawn in the last section.

II. Background

Although students know and recognise that cheating during their examinations is an ethically unacceptable behaviour, most of them admit to having done it at some point during their academic years [8]. According to Albrecht [9], three circumstances must be present for the student to be driven to cheat: some sort of pressure, the possibility of not getting caught, and the ability to rationalise the action as acceptable.

Concerning being under some sort of pressure, final examinations are, by nature, stressful and a source of anxiety for students [10]. Moreover, the personal and family concerns of living through a pandemic can exacerbate both pressure and stress [11], [12].

Lecturers are blind to what each student is doing while taking the online exam; for example, whether they are accompanied by someone who can help them in the examination, whether they are taking the examination collaboratively with other students via online media, or even whether they look up course materials that they should not consult [13]. The lack of control in online assessment encourages students who ”massively copy and plagiarise” anyway, to do it more and more often; and even more in the case of multiple-choice questionnaires, where it is easier to cheat [14].

Setting multiple-choice questionnaires based on the random selection of questions from a pool makes it more difficult to share content as fewer questions are repeated among students' exams. However, if the question pool is used repeatedly in later editions of the same examination, most of the questions become known to the latest students to take the exam. [15].

In computer programming courses, the difference between cheating and collaboration is a bit unclear [16]. Computer engineering students often undertake pair programming assignments, and this partnership, which began with pairs of students handing in assignments, may turn into fraudulent collaborations on examinations [17].

Tools are available to detect cheating by students (i.e., plagiarism detection systems); for example, Turnitin and Viper are the most widely used in higher education [18]. Although they are effective tools, they are hardly applicable to multiple-choice questionnaires and are more focused on looking for semantic similarities between sentences and between words, which is more suitable to detect unfair practice in projects or memorandums [19].

E-proctoring tools have become popular in the pandemic context and they have been used by some educational institutions to detect fraud in online examinations [20]. By using these tools, lecturers can remotely monitor students while they are taking the examination. In this research, conducted with computer engineering students, the lecturers concluded that their students cheated on online examinations, as they found significant differences between the grades of proctored and non-proctored students [21]. Unfortunately, institutions have to provide a tool to their teaching staff, who are reluctant to implement e-proctoring because of the security and privacy issues that it entails [22].

In contrast to such preventive strategies, learning analytics allow lecturers to collect evidence of the students' work in an LMS to implement pre-emptive countermeasures for cheating. Previous work has demonstrated the effectiveness of these evidence-gathering techniques when used to assess individual student performance on skills or learning outcomes [23]–[25]. In other works, researchers have collected evidence of collaboration between students performing an assignment in a group [26], [27].

Based on previous work on evidence collection from LMS activity records, this research aims to collect evidence to detect unfair collaborations that the students should not make during online examinations based on their submission timestamps and the grades obtained.

Most of the studies in the literature that evaluate using multiple-choice questionnaires focus on preventing student cheating before it occurs [28], but not on detecting it afterwards. Several of the proposals used in computer engineering courses for this aim are based on software that automatically creates customised questionnaires [29], [30]. In this way, any two students' questionnaires will be different and it will be more difficult for them to benefit from sharing their content.

In a recent work [31], researchers develop an intelligent agent that tries to anticipate fraudulent student behaviour in real-time. The agent uses both IP addresses and time of response to questions to detect suspicious patterns of behaviour.

Finally, Jaramillo et al. present an algorithm to detect students who collaborate fraudulently by sharing questions/answers. The algorithm is based on submission timestamps and exam responses [32].

III. Materials and Methods

This research proposes a method to detect students' cheating behaviour when taking examinations on LMS based on their submission timestamps and the grades obtained. This method is presented as a process model that uses the LMS activity records to identify evidence of fraudulent collaboration.

A. Model for Cheating Detection

Fig. 1 shows the model that we proposed for detecting evidence of fraudulent collaboration during students' examinations. The implementation of the model requires, first, an LMS to create examinations or assignments. Second, the LMS must allow lecturers to access and download the activity records. Finally, a software tool such as Py-Cheat is required to process the learning records and look for evidence of cheating. The model consists of a series of steps, as described below:

1. Design assessment instrument: the lecturer designs the examination on the LMS following the course syllabus.
2. Taking assessment instrument: students complete the task required in the assessment instrument.
3. Collect LMS records: the lecturer downloads the information from the LMS activity records.

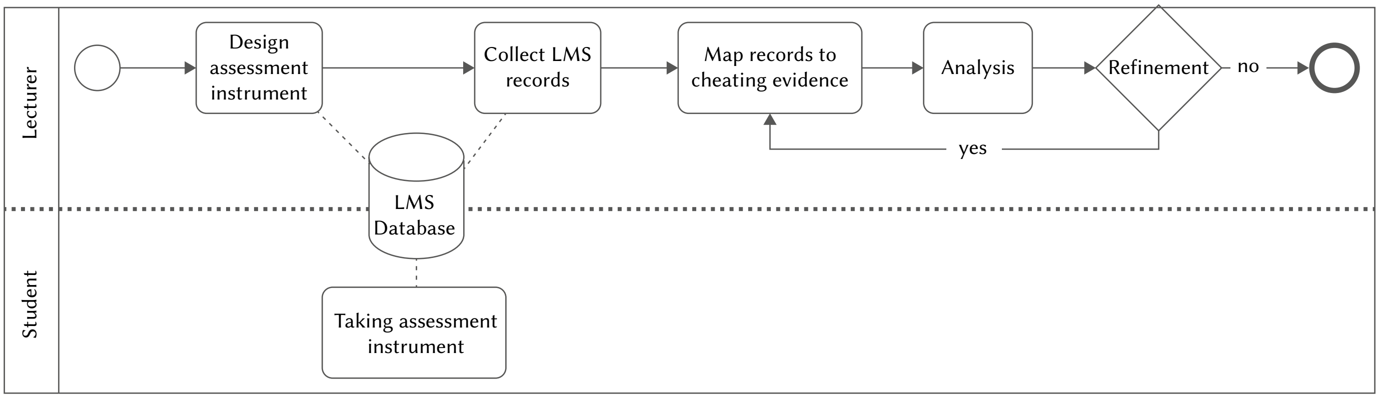
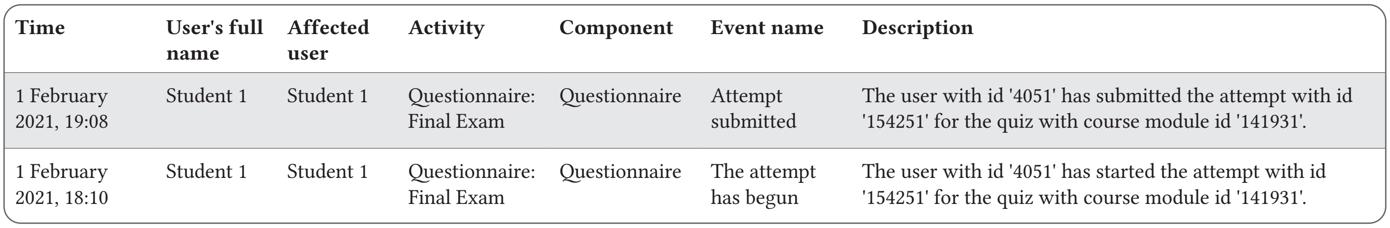

Lecturer
Student
Design assessment instrument
Collect LMS records
Map records to cheating evidence
Analysis
Refinement
no
yes
LMS Database
Taking assessment instrument

Fig. 1. Model for detecting evidence of cheating on student examinations.



Time	User's full name	Affected user	Activity	Component	Event name	Description
1 February 2021, 19:08	Student 1	Student 1	Questionnaire: Final Exam	Questionnaire	Attempt submitted	The user with id '4051' has submitted the attempt with id '154251' for the quiz with course module id '141931'.
1 February 2021, 18:10	Student 1	Student 1	Questionnaire: Final Exam	Questionnaire	The attempt has begun	The user with id '4051' has started the attempt with id '154251' for the quiz with course module id '141931'.

Fig. 2. Example of the start and end records of a student's completion of a questionnaire.

4. Map records to cheating evidence: the software tool supports the lecturer in mapping the students' records to different behaviours that can be evidence of fraudulent collaboration.
5. Analysis: the lecturer analyses the evidence and compares it with their observations of the course.
6. Refinement: the lecturer can finish the process or refine the evidence, either because they discard the previous ones or to reinforce those that were previously collected.

B. LMS Activity Records

From the information that can be found in the LMS activity records regarding student's action, this method requires the following:

- Timestamp (*T*): timestamp at which the action was carried out.
- Student (*S*): student who carried out the action.
- Activity (*A*): activity in which the student participated.
- Event (*E*): action performed by the student (e.g., access, respond, submit, etc.)

This information is enriched with the grade (*G*) obtained by the student in the examination and the laboratory group (*L*) to which the student belongs.

C. Evidence of Cheating

The method requires two events of each student's online examination completion to provide the evidence: the start time (*ST*) and the finish time (*FT*). This information is obtained from the processing of the LMS activity records. In the Fig. 2, we can see an example of the start and finish records for an activity. In this example, the activity is the ”Questionnaire: Final exam”. Its start time is February 1st, 18:10 (*ST* = 18:10 2021-02-01), taken from the event ”The attempt has begun”. Its submission time is February 1st, 19:08 (*FT* =19:08 2021-02-01), taken from the event ”Attempt submitted”. Therefore, the completion time (*T*) is 58 minutes ($T = FT - ST$).

Thus, this method returns sets of students who took the examination or who submitted the assignment sequentially and probably collaborated on it. To detect the collaboration, the method is based on the values of three features of two students' examinations. Given two students:

$S_1 \longrightarrow \{L_1, ST_1, FT_1, T_1, G_1\}$

$S_2 \longrightarrow \{L_2, ST_2, FT_2, T_2, G_2\}$

S_1 and S_2 are considered to have collaborated in carrying out an activity if:

1. S_2 starts the examination right after S_1 submits it within a time interval (*I*) defined by the lecturer: $FT_1 <= ST_2$ and $ST_2 - FT_1 = I$
2. S_2 improves the grade/minutes ratio with respect to S_1: $G_2/T_2 = G_1/T_1$ This evidence arises from two observations:
    - S_2 usually takes the same or less time to complete the examination than S_1: $T_2 <= T_1$
    - S_2 usually achieves a grade equal to or greater than S_1: $G_1 <= G_2$
3. Additionally, the lecturer can configure the results to check if the students belong to the same laboratory group: $L_1 = L_2$

Fig. 3 shows an example. If the method returns a cluster of five students (S_1, S_2, S_3, S_4 and S_5), this means that:

- Students completed the exam sequentially. While student S_1 (examinee) took the exam through the LMS, the students S_2, S_3, S_4 and, S_5 (collaborators) helped S_1 to solve it.
- Sequentially, the roles of examinee and collaborators changed until the five exams were completed.
- The last members of a sequence are likely to get higher grades. Firstly, because they have been able to repeat exam questions and, secondly, because they have had more time to search for the answers.

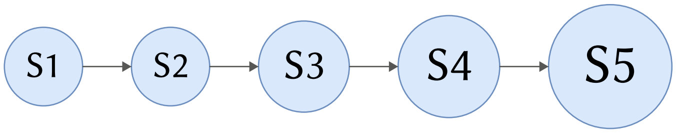


Fig. 3. Example of a cluster with five students (represented by nodes) who took the examination sequentially and helped each other in solving the exam. The diameter of the nodes symbolises the student's grade. The later they take the examination, the higher grade they get.

D. Py-Cheat Tool

Py-Cheat[1] is an open-source software tool that implements our proposal to detect students' fraudulent collaboration on activities based on the LMS activity records. The objective of Py-Cheat is to convert the collected records into evidence that can be more easily analysed by the lecturer.

Py-Cheat requires the lecturer to provide a CSV file with the following information: student (*S*), laboratory group (*L*), start time of the activity (*ST*), submission time of the activity (*FT*), completion time (*T*) and grade (*G*).

Once the lecturer provides the CSV file, Py-Cheat requests the lecturer to set the following parameters:

- Interval (*I*): maximum time that can elapse between two students submitting their assignment consecutively to be considered suspicious of fraudulent collaboration.
- Minimum number of students: the minimum number of students to check for fraudulent collaborations.

Py-Cheat processes two types of activities: assignments and questionnaires. In both cases, Py-Cheat returns a set of students who have made sequential submissions according to the specified criteria. However, there is a difference for assignments because, as opposed to questionnaire examinations, it does not take into account the duration. This happens because the start date of an examination is specifically defined in the record for each student (i.e., when the student clicks the start examination button). However, the start time of an assignment is common to all, and the LMS only records the time at which students submit the assignment. For the interval between two students in a sequence when submitting an assignment, Py-Cheat considers the difference between the submission dates: $FT_2 - FT_1 <= I$.

Finally, Py-Cheat also provides a directed graph that represents a network of students who collaborated in the examination or the assignment according to the specified criteria.

IV. Case Study

The participants are 132 students from the University of Cadiz (Spain) who were enrolled in Databases, which is a second-year compulsory course on the Computer Engineering Degree during the second semester of the 2019-20 academic semester.

In this university, the LMS is based on Moodle. We used Py-Cheat to analyse the activity records of the following activities:

- Questionnaire: a 10-multiple-choice examination corresponding to the final SQL language practice examination of the course.
- Assignments: five SQL-language practical assignments that students had to submit during the semester.

A. Questionnaire

Based on the students' complaints, as mentioned in the introduction, the examination was configured asynchronously to avoid possible problems of LMS downtime. Similarly, students were given 2 minutes and 30 seconds to answer each question, which addressed their complaint about the limited time that they had in other courses.

- A total of 10 multiple-choice questions were presented in sequence (return to previous questions was not allowed).
- Once started, the examination could not be paused.
- Asynchronous exam: 25-minute examination available for 3 hours, from 11:00 to 14:00.
- The questions are randomly selected from a pool of 100 questions. They are categorised according to their topic and level of difficulty and, for each category, there are 10 questions.

Regarding examination participation, the percentage of enrolled students who took the examination in the current 2019-20 semester was significantly higher than in the three previous semesters, for which the percentage of students who took the examination was between 62% and 67%. This rate reached 79% in the 2019-20 academic semester. A total of 105 students took the examination, out of whom 81 (77%) passed. Table I shows the examination grades in the previous four semesters. The lecturers teaching in the four semesters have been the same and, the examinations designed were of the same difficulty. Comparing the 2019-20 academic semester with the previous semesters, the grades are significantly better. For example, 59%, 71% and 45% of students failed in this examination in previous semesters with a face-to-face examination, while only 23% of students who completed the examination failed in the 2019-20 semester.

TABLE I. Grades for the Examination

Grades	2016-17	2017-18	2018-19	2019-20
A	0%	0%	3%	11%
B	9%	2%	12%	40%
C	32%	27%	40%	26%
D	59%	71%	45%	23%

B. Assignments

Throughout the 2019-20 semester, the students had to submit five assignments (A1 to A5) to the platform within a defined deadline. The assignments had the following characteristics:

- There are six lab groups, in each of which an assignment was proposed. Therefore, the assignment that each student had to submit depended on the practice group he/she was assigned.
- The students had four days to complete and submit each assignment.
- Students must perform the assignment individually.

Lecturers know that students have programmed in groups in previous programming subjects and, in many cases, they are used to working in this way. Therefore, it is likely that two students who work together beforehand will continue to work together in this subject and probably also submit the assignment simultaneously. However, the lecturers encouraged solving the assignments individually because the final assessment is individual.

Table II shows the total number of students who submitted each assignment and the total number of students who did not.

TABLE II. Total Number of Assignments Submitted and Not Submitted During the 2019-20 Academic Semester

Assignments	A1	A2	A3	A4	A5
Submitted	112	100	109	102	87
Not submitted	20	32	23	30	45

We consider the assignments to analyse whether there is a correlation between the clusters of students who can be detected taking the examination together and those clusters of students who worked together on assignment submissions.

V. Results

Once all of the students had finished the examination, the lecturer downloaded the examination records. Surprisingly, the students' completion of the examination was evenly distributed over the

[1] https://github.com/abalderas/Py-Cheat

3-hour interval that it was available (see Fig. 4). This distribution was unexpected because the students did not have any other overlapping classes. In the following subsections, we analyse the examination records through the Py-Cheat tool to find evidence of the students' fraudulent behaviour. Next, we analyse the records of the assignments carried out by the students throughout the course. In this analysis, we look for evidence of collaboration similar to those found in the examination.

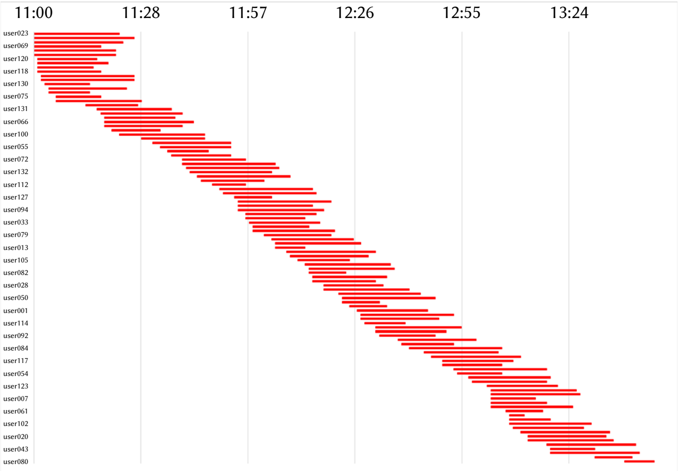


Fig. 4. Time distribution of the examination between the 3-hours slot.

A. Exam Analysis

We first draw clusters of two or more students who have sequentially taken the examination with a time interval of up to two minutes between the first student (S_1) submitting the examination and the second student (S_2) starting it; that is, $ST_2 - FT_1 <= 2min$. The results obtained indicate that 71 students met this pattern (see Fig. 5). Students assigned to the same cluster is represented by the colour of the nodes.

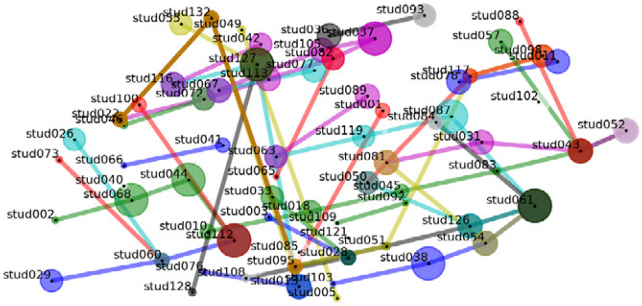


Fig. 5. Clusters of two or more students taking their examination consecutively within a 2-minute interval.

In Fig. 6, we prune the previous network by keeping only the clusters of at least five students who have taken the test consecutively with an interval of fewer than 2 minutes. In this case, four clusters of students appear. It is worth noting that some clusters overlap. Stud022 is the first to start the examination. When they finishes, stud132 and stud072 start. When stud051 finishes, stud117 and stud126 start, which would imply that there are clusters of students who worked in parallel. The diameter of the nodes represents the student's grade. The nodes have a larger diameter when they are closer to the deadline (i.e., the later they take the examination).

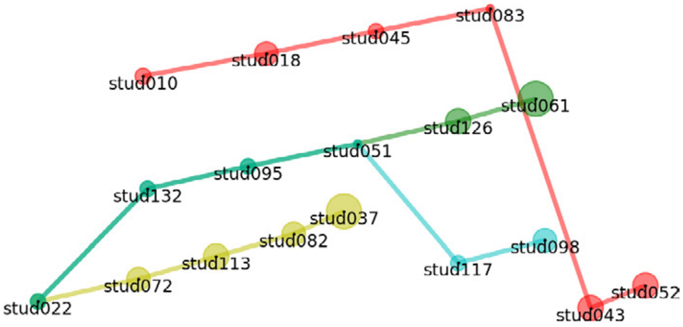


Fig. 6. Clusters of five or more students taking their examination consecutively within a 2-minutes interval.

Finally, Fig. 7 shows clusters of six or more students who consecutively took the examination within a 5-minutes interval. Table III shows the data for one cluster of students who took the examination with a time interval of fewer than 5 minutes. In this case, the first student took 25 minutes to take their examination and, from that point on, the time decreases until the sixth student takes only 10 minutes. Grades tend to be better for the last students to take the examination. Although the multiple-choice questions are randomly selected from a pool of 100 questions, there may be repetitions as new examinations are generated. Assuming that a cluster of students collaborate to take the examination, the last few students who take the examination would benefit from the repeated questions. It is worth noting that in this cluster of students, five belonged to the same laboratory group (G2).

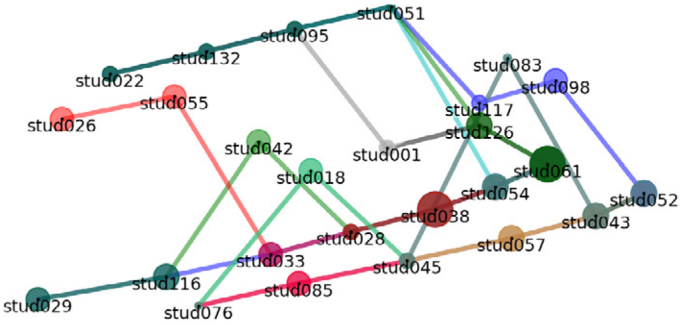


Fig. 7. Clusters of six or more students taking their examination consecutively within a 5-minutes interval.

TABLE III. Cluster of Students Sequentially Taking the Examination

Student	Lab	Start	Finish	Time	Grade	Grade /min
stud076	G3	11:40	12:05	25	3.75	0.15
stud018	G2	12:08	12:32	24	7.50	0.31
stud045	G2	12:32	12:55	23	6.25	0.27
stud057	G2	12:58	13:18	20	8.75	0.43
stud043	G2	13:19	13:31	12	8.75	0.73
stud052	G2	13:31	13:41	10	8.75	0.87

B. Assignment Analysis

Concerning assignments, we considered two possible scenarios. First, given that it is very typical to work in pairs in many courses of the degree, we checked clusters of two or more students who would submit assignments within an interval of fewer than 5 minutes. Based on the evidence collected, 97 students collaborated on some of the assignments with at least one classmate.

Figure 8 shows the clusters of students who worked together on assignment A3. Most of the clusters submitted their assignment when it had just been activated (left cluster) and when the deadline was close (right cluster). Meanwhile, only two pairs of students performed their work at an intermediate point (middle cluster).

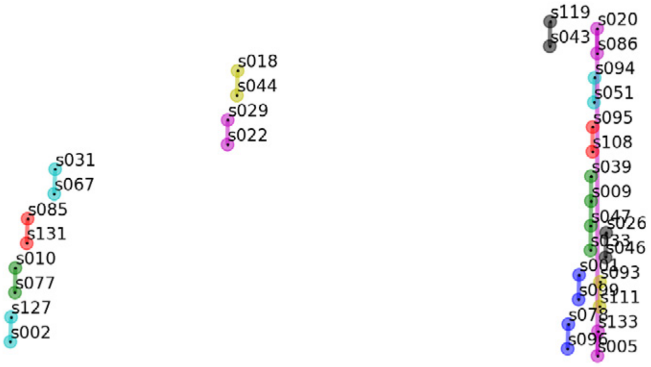


Fig. 8. Clusters of two or more students submitting their assign- ment consecutively within a 5-minutes interval for assignment A3.

The clusters of students detected in the 5-minutes interval for assignments are frequently composed of two students (Table IV). Students complete the assignment in collaboration: one student submits it first and another does it afterwards. For assignments A2, A3 and A4, Py-Cheat identifies 12, 13 and 10 pairs of students, respectively. The largest cluster of students detected was of eight students for assignment A1.

TABLE IV. Total Number of Clusters Detected for Each Assignment With the Number of Students Indicated in the Column Header: 5-Minutes Interval, Minimum of 2 Students

Student per cluster	2	3	4	5	6	7	8
A1	4	6	0	1	1	2	1
A2	12	1	1	0	0	0	0
A3	13	0	2	0	0	0	0
A4	10	0	0	0	0	0	0
A5	4	2	0	0	0	0	0

After comparing these data with those obtained in the examination (clusters of two or more students taking their examination consecutively within a 5-minutes interval), we detected that 46 students appear in collaborative clusters in both activities. In other words, we found evidence that 46 students first collaborated in the assignments and then collaborated in the multiple-choice questionnaire.

Second, the lecturer looked for large clusters of students (five or more) who coordinated to complete the assignments with an interval of fewer than 10 minutes between one submission and the next. Following this approach, 51 students participated in clusters of at least five students in the five assignments of the course (see Fig. 9).

As summarised in Table V, Py-Cheat detected large clusters of students for the first three assignments (A1, A2 and A3). Two clusters of 10 students stand out for assignment A1. However, for the last assignments of the course (A4 and A5), Py-Cheat did not detect large clusters of students.

If we compare these data with the clusters of two or more students taking their examination consecutively within a 5-minutes interval, we found evidence that 24 students first collaborated in the assignment and then collaborated in the multiple-choice questionnaire.

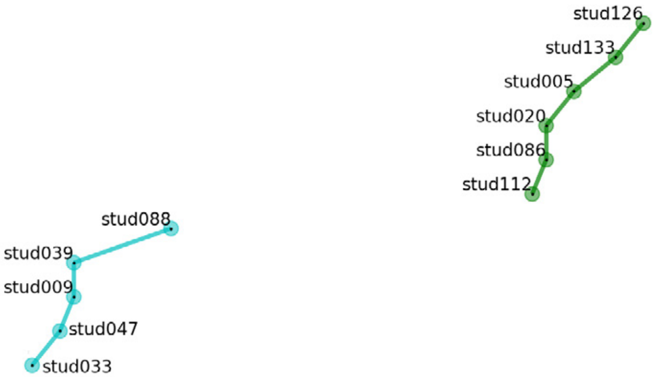


Fig. 9. Clusters of five or more students submitting their assignment consecutively within a 10-minutes interval for assignment A3.

TABLE V. Total Number of Clusters Detected for Each Assignment With the Number of Students Indicated in the Column Header: 10-Minutes interval, Minimum of Five Students

Student per cluster	5	6	7	8	9	10
A1	0	1	1	1	0	2
A2	1	0	0	0	0	0
A3	1	1	0	0	0	0
A4	0	0	0	0	0	0
A5	0	0	0	0	0	0

C. Validation

To evaluate whether the collaboration detected through the Py-Cheat application has been significant, we compare the performance of students who successfully passed the examination with students from previous years.

After the questionnaire that we used in this study, the students were required to take an examination on SQL queries. In the following, we compare the dependence relationship among the three previous courses between the students who passed this questionnaire and those who passed the SQL queries examination. We will then perform the same comparison with the students in the case study (Table VI).

TABLE VI. Results of Previous Four Academic Semesters

Semester	Stud	Activity	Passes	Fails
2016-17	115	Questionnaire	27.83%	72.17%
		SQL queries exam	26.09%	73.91%
2017-18	89	Questionnaire	20.22%	79.78%
		SQL queries exam.	29.21%	70.70%
2018-19	80	Questionnaire	51.25%	48.75%
		SQL queries exam.	33.75%	66.25%
2019-20	105	Questionnaire	66.67%	33.33%
		SQL queries exam.	34.29%	65.71%

To carry out this comparison, we define the following null hypothesis:

- H_0: Passing the questionnaire and the SQL queries examination is not related. There is no dependency between students who have passed the SQL queries examination and those who previously passed the questionnaire.

Therefore, the alternative hypothesis would be stated as follows:

- H_1: Passing the questionnaire and the SQL queries examination is related; that is, students who passed the SQL queries examination may have previously passed the questionnaire.

To determine the dependency between passing the SQL queries examination after passing the questionnaire, we used the Chi-square test with a significance level of 0.05. For the 2016-17, 2017-18 and 2018-19 academic semesters, we obtain a p-value of 13.1490, 20.1829 and 23.1090, respectively. All three values are above the significance threshold of 0.05 ($X^2 > 3.8414$), so for the previous three courses in which the examination was face-to-face, we cannot accept the null hypothesis. Thus, we assume that there is a relationship between first passing the questionnaire and then passing the SQL queries examination.

However, for the 2019-20 academic semester, we obtained a p-value of 0.7608 in the Chi-square test. This value is lower than that established for a significance value of 0.05 ($X^2 < 3.8414$). Thus, the null hypothesis cannot be discarded and this supports the fact that some students have cheated. This under-performance between this cohort of students and that of previous semesters concerning passing the second examination after having passed the first supports the view that students in the online examination cheated more than in face-to-face evaluation.

VI. Discussion

The results show evidence of cheating behaviour by students based on fraudulent collaboration among them, both for course assignments and during the online questionnaire.

Why do students cheat? As Albrecht states [9], students cheat when they know that they will not be caught cheating. The lecturer cannot see what the students are doing at home while they are taking the examination. Furthermore, they are motivated to cheat by the pressure of an official examination.

Finally, Albrecht pointed to a third reason, which is the rationalisation of the action. With Py-Cheat, evidence was obtained indicating that up to 97 students may have fraudulently collaborated in their class assignments. Therefore, if the students are familiar with collaboration during in-class assignments, then they can normalise collaborating in an examination—even if it goes again the rules. The environment is the same (online), from home they can still communicate with their classmates via videoconference, phone call or messaging while the lecturer is unaware of their actions. Consequently, they rationalise their behaviour as doing the same routine that they used for assignments.

For the assignments, the collaborations detected were mainly from pairs of students. Depending on the setting used, i.e. interval between submission of assignments and the minimum number of students per cluster, evidence has been found that between 24 and 46 students collaborated first on the assignments and then on the questionnaire. This finding is in line with the research of Hellas et al. [17]. They found that computer engineering students who had practised pair programming on assignments worked also together on the take-home examination. This work focused on the search for plagiarism based on the similarity of responses. This approach is similar to other tools such as Turnitin or Viper [18], but it also allows for detecting copy-paste patterns. To carry out this, the student had to install a plugin.

The method we present allows detecting evidence of cheating regardless of the type of activity, as it considers the time of submission, no matter if they are multiple-answer questions or a piece of code. Besides, it is transparent to the student, as no plugin is required.

The use of submission timestamps to detect cheating is in line with Tiong and Lee's research [31]. They use the submission timestamp to prevent students from cheating. Compared to our method, they aim at cheating prevention instead of detection. In this paper, different multiple-choice examinations were defined with a set of questions each. If the student answered a question in less time than expected, the system changed the following questions in his/her exam. Although the authors indicated that this was an effective system to prevent cheating, other work has shown that personalising examinations can be unfair to students [30].

Our method for detecting cheating does not depend on how the lecturer sets up the questionnaire. It considers the time of submission and the grades obtained to provide evidence of fraudulent collaboration. Therefore, it does not generate unfair situations per se, as it is transparent to the student and merely reflects what is recorded on the virtual campus.

Concerning the questionnaire, larger clusters of students took the questionnaire sequentially. While one student in the cluster takes the examination (examinee), the other students in the cluster help the former to solve the exam (collaborators). In this way, they exchange their roles and perform the examination sequentially until the last one finishes. Generally, students obtain a higher grade when their position in the sequence is closer to the end. In addition, the first students in a sequence typically take longer to complete the test than the last students. As an example, we can look at Table III, where the first student in a sequence of six took 25 minutes, while the sixth student took only 10 minutes.

These results are similar to those detected in the research of Jaramillo et al. [32]. They used submission timestamps and the responses selected by students to detect their cheating behaviour. However, their method relies on questions being repeated between exams. This is contrary to good practice in online exam design, such as using question randomisation and large question pools. The evidence based on the submission timestamps collected in our work is independent of the answers submitted by the students.

This study was made possible thanks to Py-Cheat because the tool automated the extraction of evidence of student collaboration in activities and examinations conducted through an LMS. In this study, the evidence was used to confirm suspicions of fraudulent students' behaviour. Our method avoids privacy issues caused by tools such as Proctoring or Respondus. These tools use the webcam to monitor how students take the examination and, although they can be effective to detect cheaters, this causes controversy within the educational community [22].

The aim of this paper is not to investigate the psychology of students and why they cheat. Even more, based on the results gathered in this paper, we do believe that lecturers should be encouraged to use an assessment based on continuous evaluation of their courses instead of final examinations [33], [34]. Additionally, thanks to learning analytics, we can collect learning records that can be used for evidence-based assessment [35], [36].

At the end of the course, we invited students to anonymously answer a survey in which they responded to different aspects of online teaching. We asked them what caused them the main difficulty in following the course, and most of their answers mentioned problems stemming from the situation generated by the pandemic: family problems, stress, anxiety or difficulty in concentrating.

This case study took place during the Covid-19 pandemic, which has affected the whole society, including students [37]. However, the grades achieved by the students were surprisingly higher than in previous years. For this improvement, we can find two justifications. On the one hand, the students did improve over previous years [38]. However, this would be unexpected, considering the low attendance and participation in class. On the other hand, it would be more possible for the pandemic to push some students to engage in or normalise this fraudulent behaviour during examinations.

VII. Conclusions

The Covid-19 pandemic significantly affected Spanish higher education in 2020 spring semester, forcing teaching and assessment to shift from face-to-face to online overnight. Among other issues, the assessment activities had to be conducted online, which made fraudulent activities harder to detect. This research presents Py-Cheat, a tool to detect students' fraudulent collaboration in the submission of assignments and examinations trough their LMS activity records. Specifically, the method is based on indicators as the submission timestamps and the grades obtained. The evaluation case study shows evidence that a large number of students cheated during an examination based on a multiple-choice questionnaire. The students were organised in clusters and sequentially took their examinations in collaboration.

The results collected are promising. In a virtual context, where the lecturer cannot know what the students are doing, Py-Cheat provides evidence concerning the students' behaviour and it graphically draws the different clusters of students who collaborate in the completion of assignments.

We recommend the implementation of plans to raise awareness of the ethics code and dissuade cheating. We also encourage the use of learning analytics techniques and tools such as Py-Cheat to detect fraudulent behaviour among students in the performance of assignments and examinations in the context of normality.

In our future work, this tool will focus on detecting patterns of students' collaboration and incorporating new evidence to assess whether these collaborations can lead to an improved student performance.

Acknowledgment

This research was funded by Spanish National Research Agency (AEI), through the project VISAIGLE (TIN2017-85797-R) with ERDF funds.

References

[1] D. Stuber-McEwen, P. Wiseley, S. Hoggatt, "Point, click, and cheat: Frequency and type of academic dishonesty in the virtual classroom," *Online Journal of Distance Learning Administration*, vol. 12, no. 3, 2009.

[2] F. J. García-Peñalvo, "El sistema universitario ante la covid-19: Corto, medio y largo plazo," *Universídad*, 2020, doi: https://bit.ly/2YPUeXU.

[3] C. Giovannella, M. Passarelli, D. Persico, "The effects of the covid-19 pandemic on italian learning ecosystems: The school teachers' perspective at the steady state," *Interaction Design and Architecture(s)*, vol. 45, pp. 264–286, 2020.

[4] R. Harper, T. Bretag, K. Rundle, "Detecting contract cheating: examining the role of assessment type," *Higher Education Research & Development*, vol. 40, no. 2, pp. 263–278, 2021.

[5] B. Chen, M. West, C. Zilles, "How much randomization is needed to deter collaborative cheating on asynchronous exams?," in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 2018, pp. 1–10.

[6] I. M. Sierra, M. G. Gómez, J. S. Eizaguirre, "Learning analytics for formative assessment in engineering education," *The International Journal of Engineering Education*, vol. 34, no. 3, pp. 953–967, 2018.

[7] A. Álvarez-Arana, M. Larrañaga-Olagaray, M. Villamañe-Gironés, "Mejora de los procesos de evaluación mediante analítica visual del aprendizaje," *Education in the Knowledge Society*, vol. 21, no. 9, 2020, doi: 10.14201/eks.21554.

[8] R. A. Bernardi, R. L. Metzger, R. G. S. Bruno, M. A. W. Hoogkamp, L. E. Reyes, G. H. Barnaby, "Examining the decision process of students' cheating behavior: An empirical study," *Journal of Business Ethics*, vol. 50, no. 4, pp. 397–414, 2004.

[9] W. S. Albrecht, G. W. Wernz, T. L. Williams, *et al.*, *Fraud: Bringing light to the dark side of business.* Irwin Professional Pub., 1995.

[10] E. J. Austin, D. H. Saklofske, S. M. Mastoras, "Emotional intelligence, coping and exam-related stress in canadian undergraduate students," *Australian Journal of Psychology*, vol. 62, no. 1, pp. 42–50, 2010.

[11] E. T. Baloran, "Knowledge, attitudes, anxiety, and coping strategies of students during covid-19 pandemic," *Journal of Loss and Trauma*, vol. 25, no. 8, pp. 635–642, 2020.

[12] F. J. García-Peñalvo, A. Corell, V. Abella-García, M. Grande-de Prado, "Recommendations for mandatory online assessment in higher education during the covid-19 pandemic," in *Radical Solutions for Education in a Crisis Context*, Springer, 2021, pp. 85–98.

[13] G. R. Watson, J. Sottile, "Cheating in the digital age: Do students cheat more in online courses?," *Online Journal of Distance Learning Administration*, no. 13.1, 2010.

[14] S. Kocdar, A. Karadeniz, R. Peytcheva-Forsyth, V. Stoeva, "Cheating and plagiarism in e-assessment: students' perspectives," *Open Praxis*, vol. 10, no. 3, pp. 221–235, 2018.

[15] R. W. Smith, T. Prometric, "The impact of braindump sites on item exposure and item parameter drift," in *Annual Meeting of the American Education Research Association*, 2004.

[16] T. Mason, A. Gavrilovska, D. A. Joyner, "Collaboration versus cheating: Reducing code plagiarism in an online ms computer science program," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019, pp. 1004–1010.

[17] A. Hellas, J. Leinonen, P. Ihantola, "Plagiarism in take-home exams: Help-seeking, collaboration, and systematic cheating," in *Proceedings of the 2017 ACM conference on innovation and technology in computer science education*, 2017, pp. 238–243.

[18] R. R. Naik, M. B. Landge, C. N. Mahender, "A review on plagiarism detection tools," *International Journal of Computer Applications*, vol. 125, no. 11, 2015.

[19] A. Abdi, N. Idris, R. M. Alguliyev, R. M. Aliguliyev, "Pdlk: Plagiarism detection using linguistic knowledge," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8936–8946, 2015.

[20] S. E. Eaton, K. L. Turner, "Exploring academic integrity and mental health during covid-19: Rapid review," *Journal of Contemporary Education Theory & Research (JCETR)*, vol. 4, no. 2, pp. 35–41, 2020.

[21] B. Chen, S. Azad, M. Fowler, M. West, C. Zilles, "Learning to cheat: Quantifying changes in score advantage of unproctored assessments over time," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 2020, pp. 197–206.

[22] C. S. Gonzalez-Gonzalez, A. Infante-Moro, J. C. Infante-Moro, "Implementation of e-proctoring in online teaching: A study about motivational factors," *Sustainability*, vol. 12, no. 8, p. 3488, 2020.

[23] A. Balderas, J. M. Dodero, M. Palomo-Duarte, I. Ruiz-Rube, "A domain specific language for online learning competence assessments," *International Journal of Engineering Education*, vol. 31, no. 3, pp. 851–862, 2015.

[24] M. L. Sein-Echaluce, A. Fidalgo-Blanco, J. Esteban-Escano, F. J. García-Peñalvo, M. A. Conde-González, "Using learning analytics to detect authentic leadership characteristics in engineering students," *International Journal of Engineering Education*, vol. 34, no. 3, pp. 851–864, 2018.

[25] A. Balderas, L. De-La-Fuente-Valentin, M. Ortega-Gomez, J. M. Dodero, D. Burgos, "Learning management systems activity records for students' assessment of generic skills," *IEEE access*, vol. 6, pp. 15958–15968, 2018.

[26] A. Balderas, M. Palomo-Duarte, J. M. Dodero, M. S. Ibarra-Sáiz, G. Rodríguez-Gómez, "Scalable authentic assessment of collaborative work assignments in wikis," *International Journal of Educational Technology in Higher Education*, vol. 15, no. 1, pp. 1–21, 2018.

[27] F. Riquelme, R. Munoz, R. Mac Lean, R. Villarroel, T. S. Barcelos, V. H. C. de Albuquerque, "Using multimodal learning analytics to study collaboration on discussion groups," *Universal Access in the Information Society*, vol. 18, no. 3, pp. 633–643, 2019.

[28] D. Von Gruenigen, F. B. d. A. e Souza, B. Pradarelli, A. Magid, M. Cieliebak, "Best practices in e-assessments with a special focus on cheating prevention," in *2018 IEEE Global Engineering Education Conference (EDUCON)*, 2018, pp. 893–899, IEEE.

[29] S. Manoharan, "Cheat-resistant multiple-choice examinations using personalization," *Computers & Education*, vol. 130, pp. 139–151, 2019.

[30] P. Denny, S. Manoharan, U. Speidel, G. Russello, A. Chang, "On the

fairness of multiple-variant multiple-choice examinations," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019, pp. 462–468.

[31] L. C. O. Tiong, H. J. Lee, "E-cheating prevention measures: Detection of cheating at online examinations using deep learning approach–a case study," *arXiv preprint arXiv:2101.09841*, 2021.

[32] D. Jaramillo-Morillo, J. Ruipérez-Valiente, M. F. Sarasty, G. Ramírez-Gonzalez, "Identifying and characterizing students suspected of academic dishonesty in spocs for credit through learning analytics," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1–18, 2020.

[33] K. A. Villanueva, S. A. Brown, N. P. Pitterson, D. S. Hurwitz, A. Sitomer, "Teaching evaluation practices in engineering programs: Current approaches and usefulness," *International Journal of Engineering Education*, vol. 33, no. 4, pp. 1317–1334, 2017.

[34] F. J. García-Peñalvo, A. Corell, V. Abella-García, M. Grande, "Online assessment in higher education in the time of covid-19," *Education in the Knowledge Society*, vol. 21, 2020.

[35] Á. Fidalgo-Blanco, M. L. Sein-Echaluce, F. J. García-Peñalvo, M. Á. Conde, "Using learning analytics to improve team-work assessment," *Computers in Human Behavior*, vol. 47, pp. 149–156, 2015.

[36] M. Palomo-Duarte, A. Berns, A. Balderas, J. M. Dodero, D. Camacho, "Evidence-based assessment of student performance in virtual worlds," *Sustainability*, vol. 13, no. 1, p. 244, 2021.

[37] Chegg.org, "Global student survey 2021," Chegg Inc., 2021. [Online]. Available: https://www.chegg.org/global-student-survey-2021.

[38] S. Iglesias-Pradas, Á. Hernández-García, J. Chaparro-Peláez, L. Prieto, "Emergency remote teaching and students' academic performance in higher education during the covid-19 pandemic: A case study," *Computers in Human Behavior*, p. 106713, 2021.

Antonio Balderas

Antonio Balderas received his MSc degree in computer science and his PhD degree from the University of Cadiz, Spain. He is currently with the University of Cadiz, and works as an Assistant Professor in the Department of Computer Engineering and as a Researcher with the Software Process Improvement and Formal Methods Group. He was a project manager in different Spanish IT companies. His research interests include technology-enhanced learning and creative computing.

Manuel Palomo-Duarte

Manuel Palomo-Duarte received his MSc degree in computer science from the University of Seville and his PhD degree from the University of Cadiz. He works in the Computer Science Department of the University of Cadiz as an Associate Professor. He is the author of more than 20 papers published in indexed journals and than 30 contributions to international academic conferences. His main research interests are learning technologies, serious games and collaborative web. He was a board member of Wikimedia Spain from 2012 to 2016.

Juan Antonio Caballero-Hernández

J. A. Caballero-Hernández received his MSc degree in computer science and his PhD degree from the University of Cadiz, Spain. His main research interest is focused on learning experiences based on serious games and diverse applications of process mining. Outside the academic environment, he has worked in many different positions in the IT sector, including web development, managing teams and international projects.

Mercedes Rodriguez-Garcia

Mercedes Rodriguez-Garcia is an assistant lecturer in the Department of Automation Engineering, Electronics and Computer Architecture at the University of Cádiz, Spain. She received her Ph.D. in Computer Science and Mathematics of Security from the Universitat Rovira i Virgili in 2017. Her research interests include data privacy, computer network security, and reverse engineering and secure architectures.

Juan Manuel Dodero

Juan Manuel Dodero received his degree in CS from the Polytechnic University of Madrid and his PhD degree in CS from the Carlos III University of Madrid. He was a Research and Development Engineer with Intelligent Software Components S.A. He has been a lecturer with the Carlos III University of Madrid. He is currently a Full Professor with the University of Cadiz, Spain. His main research interests include Web science, and engineering and technology-enhanced learning; fields in which he has co-authored numerous contributions in journals and research conferences.

A Systematic Literature Review of Empirical Studies on Learning Analytics in Educational Games

Ahmed Tlili[1], Maiga Chang[2,7], Jewoong Moon[3], Zhichun Liu[4], Daniel Burgos[5]*, Nian-Shing Chen[6], Kinshuk[8]

[1] Smart Learning Institute of Beijing Normal University, Beijing (China)
[2] School of Computing and Information Systems, Athabasca University, Edmonton (Canada)
[3] Florida State University, Tallahassee (USA)
[4] University of Massachusetts Dartmouth, Dartmouth (USA)
[5] Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR), 26006, Logroño, La Rioja (Spain)
[6] Department of Applied Foreign Languages, National Yunlin University of Science and Technology, 123 University Road, Section 3, Douliou, Yunlin (Taiwan)
[7] Department of M-Commerce and Multimedia Applications, Asia University, 500 Liufeng Road, Taichung City, 41354 (Taiwan)
[8] University of North Texas, 3940 N. Elm Street, G 150, Denton, TX, 76207 (USA)

Received 12 October 2020 | Accepted 13 February 2021 | Published 12 March 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

Learning analytics (LA) in educational games is considered an emerging practice due to its potential of enhancing the learning process. Growing research on formative assessment has shed light on the ways in which students' meaningful and in-situ learning experiences can be supported through educational games. To understand learners' playful experiences during gameplay, researchers have applied LA, which focuses on understanding students' in-game behaviour trajectories and personal learning needs during play. However, there is a lack of studies exploring how further research on LA in educational games can be conducted. Only a few analyses have discussed how LA has been designed, integrated, and implemented in educational games. Accordingly, this systematic literature review examined how LA in educational games has evolved. The study findings suggest that: (1) there is an increasing need to consider factors such as student modelling, iterative game design and personalisation when designing and implementing LA through educational games; and (2) the use of LA creates several challenges from technical, data management and ethical perspectives. In addition to outlining these findings, this article offers important notes for practitioners, and discusses the implications of the study's results.

Keywords

Learning Analytics, Educational Data Mining, Educational Games, Systematic Literature Review.

DOI: 10.9781/ijimai.2021.03.003

I. Introduction

EDUCATIONAL games are immersive, interactive and can engage students in dynamic learning-and-playing processes. They have therefore been used in various disciplines, such as computer architecture [1], mathematics [2], language learning [3][4] and science [5], as well as in online contexts [6][7]. Unlike in traditional learning management systems, users generate massive data while interacting with educational games. Therefore, collecting and synthesising students' en-route behaviour patterns, intellectual states and emotional level in gameplay become essential to identifying how their playful learning occurs. Researchers have sought various ways to utilise this data to identify how to accurately observe students' learning process through learning analytics (LA) [8]. LA refers to the collection and analysis of learners' intellectual and behavioural attributes to optimise learning experiences [9]. Several studies on educational games have also focused on adopting unobtrusive ways to use LA approaches to measure students' progressions without interrupting the flow of their gameplay [10] [11]. In particular, using stealth assessment in educational games has broadened the role of real-time and automatic assessments in educational games [12]. Moving away from existing approaches that rely on external measures (e.g. post-test), recent research has sought ways to promptly measure how students' in-situ learning occurs while they are experiencing ongoing gameplay.

The LA field has expanded in recent years because it allows educators to perform formative assessments that accompany fine-grained and contextual feedback tailored to the various needs of individuals in learning environments. For instance, recent implementations of educational games have integrated formative assessments with LA [13] [14]. Serving as a mechanism of such assessment, LA in educational games aims to identify and interpret students' meaningful progressions or task challenges in gameplay. Subsequently, game elements and supports (e.g. feedback, learning sequence and presentation of materials) are tailored to individual needs (e.g. domain knowledge, cognitive competence, or affective states).

* Corresponding author.

E-mail address: daniel.burgos@unir.net

Despite the emerging significance of LA when implementing educational games, a limitation also remains. There is a lack of comprehensive guidelines for LA design, development, and implementation, because analytic approaches are game- and context-specific, resulting in high variations in adoption. Specifically, this issue limits developers' ability to define general analytics to effectively incorporate LA in educational games [15]. In other words, the applications of LA in educational games still appear complex, and generally acceptable approaches have rarely been reported [16]. This fact implies that it is necessary to perform a comprehensive review to explore how LA has been integrated and implemented in educational games.

Accordingly, this study conducts a systematic literature review to better understand the potential implementations of educational games across various contexts. The goal of this study is to advance this field by (1) exploring why and how LA has been implemented across various learning contexts and (2) discussing existing limitations and challenges in integrating LA in educational games. This study is structured as follows: Section II presents the background of LA in educational games and highlights the research gap this study aims to address. Section III presents the research method followed to conduct the systematic literature review. Section IV presents the findings of this study, while section V discusses these findings. Finally, section VI concludes the study with general notes and future directions.

II. Background of LA in Educational Games

LA is an interdisciplinary field associated with many domains, including data science, artificial intelligence, practices of recommender systems, and online marketing and business intelligence [17]. LA is defined as 'the measurement, collection, analysis and reporting of data about learners and their context, for purposes of understanding and optimising learning and the environments in which it occurs' [18]. Powell and MacNeill [19] highlighted key applications of LA, namely to: (1) offer feedback for students about their learning performance; (2) predict at-risk students who may fail to pass their final exams; (3) help educators to provide interventions when needed; (4) improve the design of courses; and (5) support decision-making when it comes to administrative tasks. LA has been increasingly prominent because it enables researchers to collect, interpret and share meaningful data that inform how learners interact with a learning environment.

The applications of LA are rooted in the usage of formative assessment in learning. A formative assessment is one that is integrated into the learning experience without interruptions during students' gameplay [11]. Research suggests the importance of formative assessments in informing educators about which cognitive and emotional challenges students may experience. In addition, formative assessment can prompt educators to decide on the types of adaptive learning supports (in-game help and game tasks) to use to foster students' deep learning [12]. Specifically, a current stream of a stealth assessment has explored feasible implementations of formative assessment in educational games [21][22]. Since educational games yield highly interactive and massive traces of learners' in-game behaviours, researchers have considered the latent uses of LA in educational games. In the same vein, Alonso-Fernandez, Calvo, Freire, Martinez-Ortiz and Fernandez-Manjon [23], as well as Tlili and Chang [24], have stated that educational games without analytics are like black boxes that barely offer meaningful clues to students' learning process during their play.

Hence, LA is applied in educational games to better capture how students' improvement and challenges occur without interrupting their flow, and then to inform tailored feedback for game-based learning experiences. However, previous studies rarely suggested how and why LA techniques are capable of supporting learners' play in educational games. This gap demonstrates that it is essential to understand (1) what are the objectives of implementing LA in educational games; (2) what are the educational game contexts; and (3) how such factors (objectives and game contexts) can influence various LA implementations in educational games.

Despite the potential of future combinations of LA and educational games, integrating those two systems remains challenging. Papamitsiou and Economides [25] asserted that further explorations using LA in educational games are necessary because understandings of the intersection between LA and interactive learning environments are still vague. Saveski et al. [26] revealed that 21 European game studies demonstrated a high interest in applying LA in educational games, but the researchers were concerned with the complexity of implementation. Like previous researchers, Perez-Colado, Perez-Colado, Freire-Moran, Martinez-Ortiz and Fernandez-Manjon [16] mentioned that the application of LA in educational games is still a complicated process, despite the fact that there are several platforms which combine both educational games and analytics. Therefore, given the gap between the advancement of LA technologies and their practical implementations in educational games, a further systematic literature review is necessary to gain insights that can close the gap. To address the questions above, we proposed four primary research questions in this study.

- RQ1. What are the objectives of applying LA in educational games?
- RQ2. What genres of educational games have applied LA, and what types of game metrics were used in the application of LA?
- RQ3. What types of LA approaches were used in educational games?
- RQ4. What are the challenges in applying LA in educational games?

III. Method

A systematic literature review of empirical studies using LA in educational games was conducted based on the major steps outlined by Okoli and Schabram [27].

A. Data Collection and Search Criteria

Several keywords, including 'learning analytics AND educational games', 'learning analytics AND game-based learning' and 'educational data mining in games' were used in searches in different electronic databases, namely Taylor & Francis Online, IEEE Xplore Digital Library, ScienceDirect, AIS Electronic Library, Springer, Wiley Online Library, ACM Digital Library, ProQuest and Semantic scholar. As shown in Fig. 1, these searches yielded a total of 405 studies conducted from 2012 to 2019. Of those, 180 studies were removed since they were found to be duplicated. The remaining 225 studies were then evaluated by title, abstract and, if necessary, by full text, based on the inclusion and exclusion criteria described in Table I. In the end, only 36 studies met the inclusion criteria, and those studies were double-checked again through readings of the full text.

TABLE I. Inclusion and Exclusion Criteria

Inclusion and Exclusion Criteria technique
Articles that are published in peer-reviewed journals.
Articles that are empirical studies and report rigorous study procedures and their findings.
Articles that involve human subjects.
Articles that have their full text available online.
Articles that apply LA in educational games.

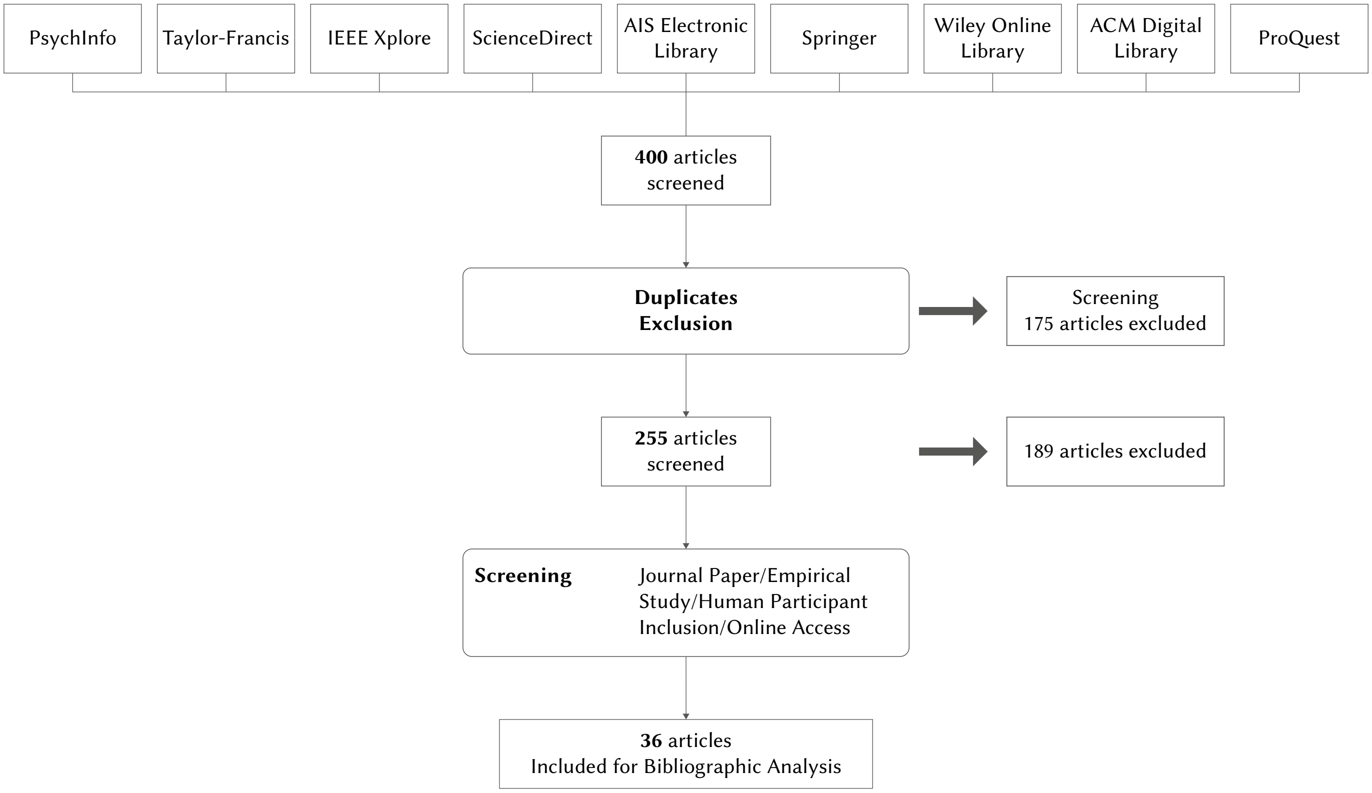

PsychInfo
Taylor-Francis
IEEE Xplore
ScienceDirect
AIS Electronic Library
Springer
Wiley Online Library
ACM Digital Library
ProQuest
400 articles screened
Duplicates Exclusion
Screening 175 articles excluded
255 articles screened
189 articles excluded
Screening
Journal Paper/Empirical Study/Human Participant Inclusion/Online Access
36 articles Included for Bibliographic Analysis

Fig. 1. Flowchart of the systematic review process.

B. Coding Procedure

Each study in the literature was coded for different characteristics. As this study aimed to explore LA designs and implementations in educational games, we took into consideration existing integrations of LA in combination with major features of educational games discussed in the sampled studies. Hence, two major coding schemes presented in two systematic literature reviews guided our coding scheme in this study, namely Papamitsiou and Economides [25] for LA and Connolly et al. [28] for educational games. In addition to the basic article information (e.g. year of the study, journal name), we evaluated the LA objective and technical information (i.e. LA approach, whether the LA was embedded in the game or not and LA challenges). In addition, we evaluated environmental information (i.e. game genre and game mode) and the metrics of the educational games highlighted in these studies. Table II presents the detailed coding scheme of this study. Finally, suggested by Webster and Watson [29], the coding results were then organized in a table (see Appendix I), which formed the abstract of this systematic literature review.

We designed an initial coding protocol for multiple coders who are experts on educational games and LA. Based on the coding scheme presented in Table II, we carried out training using a subset of the literature for this study. The coder training was conducted until all coders reached consensus. When individuals' coding results differed, all coders iteratively discussed their results to reach an agreement. A detailed summary of the articles reviewed based on the coding variables is presented in Appendix I.

TABLE II. Coding of Reviewed Research Papers Examining LA in Educational Games

Variables	Description	Coding Criteria
Year	Year study was conducted	Year
LA objective	Goal of applying LA in educational games	• Understanding and modelling students' in-game behaviours • Formative design of educational games • Implementing teaching supports • Conducting learning assessments • In-game personalisation
LA approach	The approach used to analyse data	• Data mining and analytics (e.g. lag-sequential analysis and social network analysis) • Data visualization • Sequential data analytics
Challenge	The challenges of LA application in educational games	Report the mentioned challenges during the application of LA in educational games
Embedded analytics	Was the LA procedure incorporated within the educational game?	Yes / No
Game mode	The mode of the educational game used while applying LA	Single player / multi-player / massively multiplayer
Game metrics	The metrics used within educational game for LA	Report the actual collected traces for LA

IV. Findings

A. RQ1. What Are the Objectives of Applying LA in Educational Games?

This section found major objectives of LA in educational games: understanding and modelling students' in-game behaviours (13 studies); creating formative designs for educational games (7 studies); implementing teaching support (8 studies); conducting learning assessments (13 studies); personalising in-game features (2 studies). It should be noted that several studies applied LA in educational games for more than one purpose [30]. Each key application is described in the subsequent sections. Appendix I includes the details of coded articles for this study.

1. Understanding and Modelling Students' In-game Behaviours

A group of studies has examined students' in-game behaviours [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] to identify gameplay patterns. Using collected data about such behaviours, researchers aimed at understanding students' behaviour patterns and their inherited characteristics during gameplay. For example:

- [40] Liu et al. used LA to implement in-game behaviour log analysis to collect students' time- and date-stamped actions. They implemented data visualisations of collected student log data to represent students' pattern of use of in-game tools (e.g. database, notebook, and probes).
- [42] Martin et al. used LA to explore how groups of students solved in-game tasks related to fractions. Using hierarchical clustering, this study identified three gameplay patterns.

Furthermore, researchers also observed students' in-game behaviours to obtain evidence of modelling that reflected either affective or cognitive states attained during gameplay.

- Affective state: Denden et al [33] and Essalmi et al [34] modelled the students' personalities, specifically extraversion and openness dimensions, based on their gaming behaviours and LA [9, 10].
- Cognitive state: Khenissi et al. [39] implemented LA in a memory match game to implicitly model the students' working memory capacity using a fuzzy logic algorithm.

2. Formative Design of Educational Games

Using LA was also beneficial in carrying out formative designs of educational games [15] [44] [45] [46] [47] [48] [49]. Specifically, LA enabled game designers and educators to understand how educational games can be better designed through assessments of students' play logs and observed behaviour. As examples of such design studies on educational games:

- [15] Serrano-Laguna, Torrente, Moreno-Ger and Fernández-Manjón carried out data visualisations to detect cases of outlier behaviour from students. Using collected data, they identified the types of learner challenges that appeared during gameplay.
- [49] Chaudy and Connolly used the LA engine EngAGe integrated with an educational game. This engine is designed to allow both game developers and educators to conduct iterative designs of an assessment that is capable of in-game adaptions to players.

3. Implementing Teaching Supports

LA played a role in the development of teaching supports to foster students' learning in educational games [30] [45] [46] [50] [51] [52] [53] [54]. First, researchers used LA to provide students' records of learning profiles in a game system. The use of teaching supports was found to enhance students' attention and then help them cope with challenges during gameplay. Some studies included examples of how teaching supports were applied and implemented across different domains.

- Providing a visual dashboard demonstrating real-time behaviour data
  - [30] Minović, Milovanović, Šošević and González incorporated an analytical tool in an educational game that generated a real-time dashboard (e.g. using circular graphs) for teachers to use when teaching computer networks. They used this dashboard to keep track of their students' trajectories and support their learning when needed.
  - [51] Chen and Lee applied LA to help students learn English vocabulary in an educational game. The game in this study tracked students' answers to inform teachers of students' learning states, providing warning messages and suggestions to enhance the learning process.
- Providing learner profile data for teachers' decision-making
  - [54] Rodríguez-Cerezo, Sarasa-Cabezuelo, Gómez-Albarrán and Sierra created an analytics tool in generated educational games for teaching computer language implementation to help teachers control their students while learning occurred and to assess their performance.

4. Conducting Learning Assessments

LA in educational games served as learning assessments. The key to assessment in educational games was to unobtrusively measure students' learning progressions across various subjects. A collection of studies implemented LA to identify learners' progression in in-game performance, problem-solving skills or knowledge acquisition [11] [15] [30] [52] [55] [56] [57] [58] [59] [60] [61] [62] [63].

- In-game performance [58][59][60]: a group of researchers used LA in an educational game to model students' knowledge levels, in order to allow researchers to compare expert and novice scores. Such studies used similarity indices that represented to what extent students' in-game performance emerged.
- Problem-solving [56]: Hernández-Lara, Perera-Lluna and Serradell-López applied LA in a simulation game that taught students decision-making and management skills. They aimed to implicitly observe students' interaction behaviours in relation to target learning outcomes.
- Knowledge [61]: Rowe et al. considered LA approaches in two developed educational games (Impulse and Quantum Spectre). Using game log data, this study detected learners' strategic behaviours concerning various scientific concepts (e.g. Newtonian physics).

5. In-game Personalisation

Research has implemented personalisation to automatically provide students with adaptive learning experiences in educational games [11] [53]. Adaptivity in educational games refers to providing appropriate level of challenge and tailored feedback in an educational game [5]. In accordance with this rationale, some studies have adjusted game designs to reflect learners' needs and challenges.

- [11] Reese, Tabachnick and Kosko adopted their actionable measurement system to indicate learners' progress on their in-game performance. In their educational game CyGaMEs, the game system provided students with embedded learning support and a performance progression bar showing personalised data visualisations that indicated how close students were to meeting their in-game goals.
- [53] Kiili, Moeller and Ninaus applied LA in an educational game for teaching fractions and decimals to provide personalised hints based on each student's misconceptions.

B. RQ2. What Genres of Educational Games Have Applied LA, and What Types of Game Metrics Were Used in the Application of LA?

As Fig. 2(a) shows, the educational game genres in which LA is most often applied are role-playing games (11 studies) and puzzle games (9 studies). This finding suggests that research tend to use role-playing games because those games enable students to experience playful learning in correspondence to a given narrative in a virtual environment. Moreover, role-playing games provide a media-rich environment, including interactions, activities, and places; hence, behaviour traces can be generated, tracked and used in LA everywhere in such a game environment. Puzzle games have also been used extensively, primarily because they can facilitate students' problem-solving and reasoning [64] [65]. On the other hand, as Fig. 2(b) shows, the most used educational game type in which LA is applied is the single-player game (31 studies out of 33). This result suggests that previous games were designed to promote individuals' self-regulated actions in individual adventures.

Prior studies have used several types of in-game metrics across different game genres. Such games have applied LA to meet different

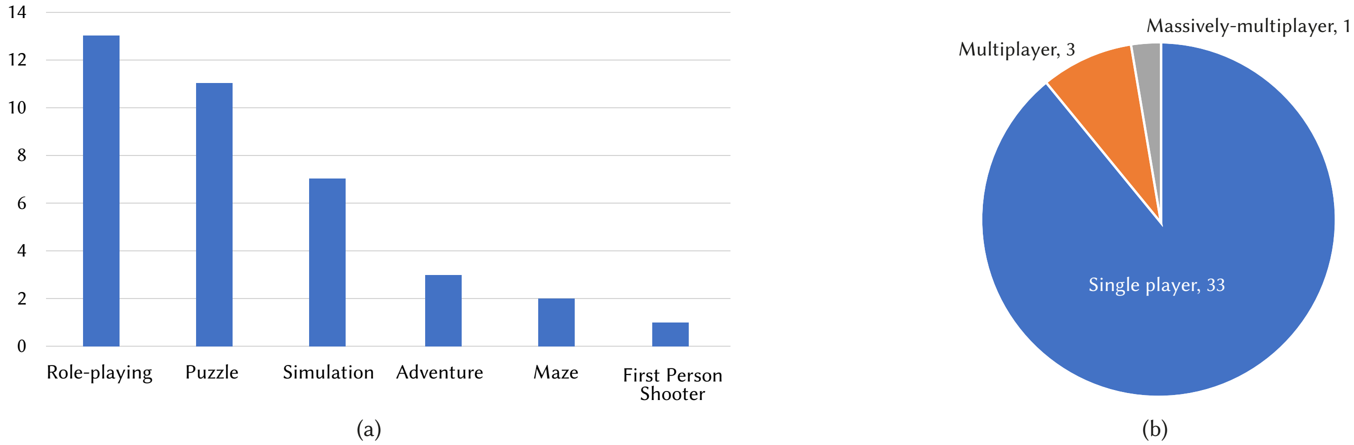


Fig. 2. Used game genres and types for la (a= game genres, b= game types).

TABLE III. Coding of Reviewed Research Papers Examining LA in Educational Games

Target measures	Type of game metrics	Examples	Description
Performance measure	In-game performance	• Game score • Reached game level • Number of correct/wrong answers	In-game performance measures can keep track of students' in-game performance in relation to learning.
	Time on task	• Time spent in each scene • Interaction time • Time solving a level	Researchers focused on measuring the time duration by either students' performance in the game in general or in a particular in-game activity or quest. This metric was specifically used to measure either how much time they paid attention to gameplay or how they efficiently accomplish game tasks.
Behavioural measures	Game interaction	• Used game characters • Interacting with game tools/ elements • Number of clicks	In terms of interaction with the game tools/elements, research focused on using the interaction of students with different game elements (found in the game environment) or game tools (provided by the game as tools to further support the learning process). The purpose of game interaction is to capture all the interactions of players.
	Learning behaviour	• Number of times using the help • Note-taking	Learning behaviours refer to the specific game interactions that are identified to be related to learning. The purpose of learning behaviour is only to capture the meaningful interactions of players.
	Progression	• Game location • Followed path • Progress in the game	These metrics focus on tracking the students' game trajectories or paths while learning in the game environment.
	Timestamp		Timestamping usually works with the game interactions, learning behaviours and progression to mark the chronological sequence of gameplay.
Multi-faceted measure	Discourse	• Dialogue • Verbal communication	Researchers focused on collecting chat/forum communications among students generated while those students were playing an educational game.
	Player information	• Personal information • Student name and age	These metrics are out of the game but can provide background information on players.
Challenge	Physical behaviour	Gesture	This type of metric looks beyond in-game performance. Bioinformation can be tracked similarly.

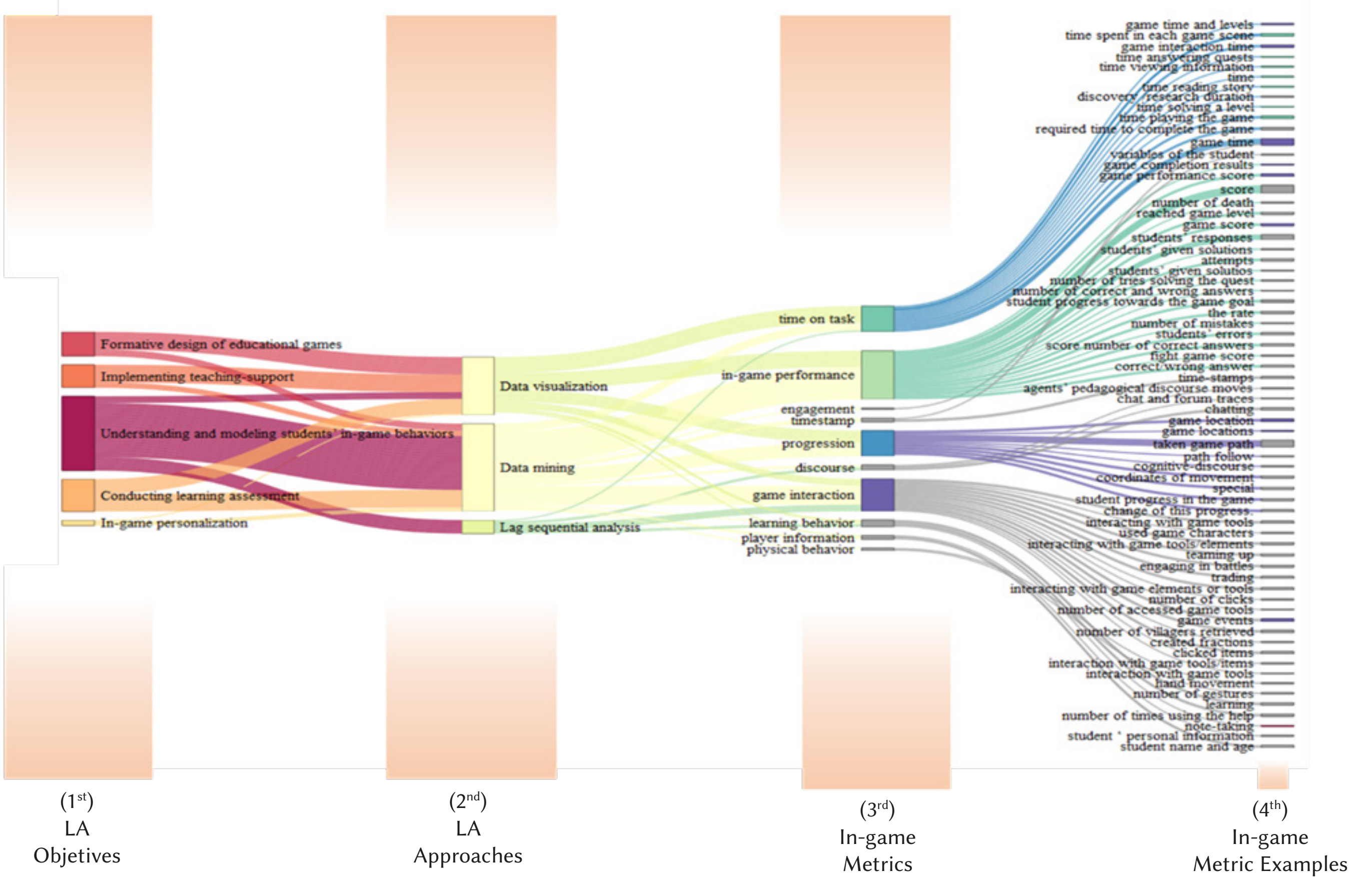


Fig. 3. Sankey diagram of LA objectives, approaches, and in-game metrics.

objectives (identified by RQ 1). Table III shows the major types of in-game metrics with their examples. In general, performance has been the most frequently emphasised target measure in previous studies. Previous studies measured either in-game performance or time on task. Another major type of measure was behavioural (e.g. game interactions, learning behaviours and progression). These metrics provided researchers with analytical information that represented 'en-route' variables of students' behaviours. Lastly, a group of LA metrics – such as discourse, player information and physical behaviour – helped researchers understand students' learning behaviours through multiple facets (e.g. discourses and physical behaviours). This type of measure proved beneficial in corroborating the results of both performance and behavioural measures but was shown to be highly dependent on inherited game contexts and data availability.

C. RQ3. What Types of LA Approaches Were Applied and Used in Educational Games?

As shown Appendix I, the most used analytics approach was data mining (28 studies), followed by data visualisation (13 studies) and sequential data analytics, or SDA (i.e. lag-sequential analysis, 2 studies). We also confirmed that several studies applied multiple analytics approaches in combination (e.g. both data mining and data visualisation). Data mining aims to discover hidden information and meaningful patterns from massive data, while SDA captures sequential transitions in behaviour events representing learning paths in gameplay [66]. Data visualisation is used to display a variety of visual stimuli, such as pie charts and histograms, to represent data indicating students' learning progression.

Most (60%) studies applied LA to educational games offline. Specifically, these studies collected the students' traces during the learning process first. They then used external data-mining software, such as Weka, to analyse these traces and extract meaningful information. The rest of the studies (40%), on the other hand, incorporated LA within the educational games to provide real-time reports for stakeholders, such as teachers and students. This result is explained by the existing limitation that designing educational games with incorporated analytics systems (i.e. automatic data collection and adaptations) is likely to be more complex and challenging. This is because to provide an adaptive game scenario, game designers and developers need to adapt all the involved game elements (e.g. mechanics, graphics, sounds, etc.) in that scenario to different profiles, which could be time consuming and with high cost. For instance, Denden et al. [33] highlighted that to provide an adaptive educational game based on personality, the game environments that a student can visit, as well the non-player characters to interact with should be personalized.

Fig. 3 is a visual synthesis of our study findings, which mapped the relationship among LA objectives, LA approaches, and the specific in-game metrics from the collected studies. It outlines a visual path indicating how LA objectives, approaches, outcome variables, and in-game metrics are interconnected. Across types of LA objectives in educational games (1st layer, 5 categories), researchers have adopted LA approaches (2nd layer, 3 categories) by drawing on a list of in-game metrics (3rd layer, 10 categories). To evaluate such outcome variables, research has used a collection of in-game metrics (4th layer, 67 examples) in the diagram. The dominant goal of LA in previous studies was to understand and model students' in-game behaviours. Accordingly, the majority of the collected studies have used data-mining techniques (2nd layer) to track students' learning as indicated by in-game performance, progression and game interactions (3rd layer). Surprisingly, few studies assessed students' outcomes in terms of learning behaviours. Most tended to focus on capturing students' in-game trajectories in relation to game performance. Our study findings suggest that to date, research on LA in educational games has mostly focused on the growth of data-mining techniques to capture, collect and explicate learners' in-game actions.

D. RQ4. What Are the Challenges in Applying LA in Educational Games?

After identifying ways of incorporating LA in educational games, this section discusses the reported challenges which can hinder such incorporation. The identified challenges (Appendix I) can be grouped into three categories: (1) techniques; (2) data; and (3) ethics.

1. Challenges on Techniques

These concerns are related to the provided infrastructure, including tools for the application of LA in educational games. They are as follows:

- Validations of learning analytics implementations in educational games:
  - Previous research also stated the difficulty of incorporating analytic systems in educational games, due to the complexity of designing educational games that contain such systems [48].
  - Fine-tuning learning analytics systems across different educational game contexts is necessary to validate the feasible adoptions of the systems (e.g. configuring either universal or contextual sets of variables and tracing rules).
- Lack of game environments to capture students' collaboration:
  - Few studies in the sampled literature configured students' collaborations during gameplay [36] [55] [56].
  - A lack of collaboration settings in previous studies suggests that there are likely limits on providing social learning experiences through group interactions (e.g. how students collaborate and what type of communication occurs). Further investigation about the application of LA in other game types, namely multiplayer and massively multiplayer, is suggested to capture the effect of social dynamics on game-based and playful learning.
- Reusability of the analytics system: According to existing studies, integrations of LA into educational games appear not to be reusable, because a high degree of variation between educational game developers and educators exists when game metrics are used [15]. Consequently, the cost of designing educational games with analytics systems is high. Therefore, researchers need to focus on developing and providing standardised and scalable LA approaches that can be used across different educational games.
- Big data storage: Educational games encourage learners to be interactive, so that their many traces can be generated and stored. This raises a question of how massive trace data from students can be stored securely [15] [54] [63]. Hence, game developers need to consider ways to store all generated trace data while avoiding losses due to network constraints.

2. Challenges in Data Collection

- Identification of important data: Research suggests that the failure to select the important data to be collected from big data storage is likely to limit the application of LA by resulting in the collection of data that cannot be useful later on. Important data should reflect, for instance, students' performance [38] [61]. In this context, Tlili et al. [67] mentioned that data generated during the application of LA should be carefully studied and selected before the collection process begins.
- Identifying relationships between data: Research states that in educational games, it is difficult to see the relationship between traces in order to extract useful information [38]. Therefore, specific metrics should be predefined (depending on the LA goal) and considered during the educational game design. This can facilitate identifying the relationships between metrics that have been previously defined.

3. Challenges in Ethics

These challenges relate to the duties and obligations that arise when applying LA in educational games.

- *Students' privacy*: Existing studies rarely consider how students' privacy is secured when collecting student data in educational games [33]. Pardo and Siemens [68] have suggested that it is necessary to consider appropriate legislation methods for data collection to avoid unintentional violations of students' privacy.
- *Transparency of LA*: Researchers have highlighted the need to ensure the transparency of collected data, which should be fully open for students. This statement implies that students should be able to retrieve their performance results when they want to during gameplay [33]. This approach can help students feel safe while applying LA and assist in immediately monitoring their learning progress by enabling a 'watch the watcher' process [69].
- *Storage time*: Researchers were not clearly aware of how long the collected data should be securely stored [33].
- *Equity challenge:* Only a few studies suggested how LA in educational games could be used to support students with disabilities. The study findings outlined in Appendix I show that 91% of educational games with LA were aimed at typical-developing students without any disabilities.

V. Discussion

A. Academic Implications

We found that existing LA practices tend to be exploratory (e.g. cluster analysis and sequential analysis). LA implementations in educational games were intended to identify learners' behaviour patterns, as well as their characteristics. Some studies employed cluster analysis to identify either individual or group learning styles and behavioural patterns [32] [38] [42]. The others used SDA to extract and visualise a major set of behaviour sequences representing strategic and engaged gameplay patterns [40] [43]. Such exploratory LA implementations were intended as a means of qualitatively analysing students' gameplay contexts and behaviour patterns through a quantitative lens.

Despite the benefits of exploratory LA studies outlined above, a challenge also exists. Existing exploratory LA practices are post-hoc analyses, which focus on showing what happens during gameplay, and are therefore limited in predicting students' learning challenges in different gameplay paths [39]. In other words, the exploratory post-hoc analyses have limited implications without validating the model performance in context. For example, existing LA practices scarcely address the question of how to provide adaptive supports to assist students in meeting challenges in educational games. Predicting learners' difficulties or challenging experiences is essential to choosing adaptive supports tailored to learners' progressions in gameplay. Hence, once an educational game aims at providing personalised game experiences to students having learning challenges, building a pipeline to connect an adaptive system with such exploratory LA can be suggested in future research.

1. Validations of Learning Analytics Measures

When using LA in educational games, only a few researchers implemented validations of various in-game data with external assessments [11] [61]. In addition to observing learners' trajectories unobtrusively, researchers should also aim to ensure that such game metrics represent target outcome variables as intended. Educational games involve multiple types of game metrics. We can build a bigger picture through multiple learner traces from gameplay. This will enable researchers to implement finer-grained data analyses (e.g., microactions), triangulate the data collected and understand learning processes in detail. For example, synthesising and validating multiple data sources from learners' gameplay is necessary to confirm how such game metrics consistently indicate learners' achievement through gameplay. Especially considering different game design contexts, validating different types of game metrics can be useful in capturing learners' meaningful gameplay reliably.

2. Lack of Learning Analytics Implementations in Collaborative Educational Games

We confirm that there is a lack of multiplayer and massively multiplayer games applying LA. This result indicates not only the necessity of designing games that foster collaboration, but also that of implementing LA to understand collaborative experiences. Although the field of LA includes various data-mining and analytics approaches used to understand learners' social dynamics, existing educational games have rarely focused on students' social interactions. Traditional ways of understanding collaboration and peer interactions among learners (e.g. observation, interview, video analysis) are generally time-consuming. LA, however, can distill emerging information much faster, sometimes with greater and unbiased detail, than those methods. It thus raises the question of how future LA practices can be contextualised in educational games that require learners' social interactions.

3. The Need for Mastery Learning Design in Educational Games

Despite the increasing growth of data-mining techniques in educational games, there is a dearth of empirical research on how to design and implement an educational game system that supports students' mastery of learning experiences. We have confirmed that previous studies largely focused on understanding and modelling students' in-game behaviours instead of on learning assessment and in-game personalisation. While the former LA objective highlights unobtrusive and externalised data collection related to students' in-game behaviours, the latter influences how likely an educational game is to be designed to enable students' mastery of learning objectives [70]. Specifically, learning assessments and in-game personalisation are means to indicate students' learning progressions and provide automatic and responsive feedback tailored to their learning states. The reported challenges in data collection are also related to the discrepancy between current and potential LA practices. The automation of learning assessments and personalisation requires an educational game to select and define important data features in relation to students' cognitive, behavioural and emotional states. However, limitations in choosing and embedding important data features into educational games and their assessment system still exist [38] [61]. This recalls our study result that few studies on educational games have focused on students' learning behaviours as a major measure in LA. Future educational games should consider ways to reinforce students' mastery of learning experiences.

B. Practical Implications

In terms of practical implications, we suggest important means through which stakeholders can understand this research field and further work better. First, a collection of game metrics (e.g. interaction with the game tools/elements, followed game path or trajectory, game time, game score, number of wrong/correct answers and chat/forum communication) has been used for different LA applications in educational games. Future educational games can use these and other metrics for different purposes, including to help researchers collect evidence that identifies learners' status in different domains (e.g. cognitive, affective and behavioural) and across various game contexts.

Second, we suggest that educators and practitioners further investigate applications of LA in educational games. Specifically, LA techniques benefit educators by enabling data-driven decisions in communication among all stakeholders. Although educational games have been increasingly used by educators in K-12 settings [5] [42], the integration of LA into educational games is still at the early stage. Such integration should foster communications between teachers and students. Besides, future LA in educational games could take more stakeholders (e.g. parents and administrators) into consideration.

Third, it is important to address the accessibility of educational games with LA. Research has shown the possibility of using LA to help students with disabilities [15] [46] [57]. However, relevant educational policies and acts, as well as inclusive game design standards are scarce. LA can help create a novel way of facilitating access to education by students with disabilities, which will further increase equity and support inclusive learning. In addition, policymakers should emphasise and address the ethical challenges of using LA. Privacy and transparency have been issues not only in educational games with LA, but also in adopting LA into educational systems.

VI. Conclusion, Limitations, and Future Work

In this study, we systematically reviewed educational games studies with LA to investigate how LA implementation has evolved. The study findings suggest that: (1) LA in educational games has been used for different purposes, such as student modelling, iterative game design, providing teaching supports and personalisation. (2) Role-playing games and puzzle games in single-player mode are the most common game-setting implementing LA. When LA has been implemented, various game metrics (e.g. interaction with the game tools/elements, followed game path or trajectory, game time, game score) have been used for data input. (3) The most frequently used analytics approaches include data mining and data visualisation. We confirmed that most of the LA approaches are post-hoc and focus on exploring students' in-game trajectories. (4) It is important to address challenges from three perspectives, namely techniques, data and ethics, to ensure the successful integration of LA applications into educational games.

It should be noted that this study has also some limitations that should be acknowledged and further researched. This study included only study findings from empirical journal articles. The narrow scope of data collection in this study limited the number of sampled data and failed to address emerging LA practices expansively. Despite such limitations, this study provides a solid basis for better understanding the ways in which LA in educational games has been contextualised. Future research could investigate ways to integrate LA solutions in educational games by providing a variety of LA examples in educational game contexts.

Appendix

APPENDIX I. A List of Selected Literatures As to Learning Analytics in Educational Games

Num	LA Goal [a]	LA Approach [b]	Game Genre [c]	Game Mode [d]	Embedded Analytics [e]	Challenges [g]	Game Metrics
[11]	Ca, Ip	Dm	Pz	Sg	No	N/A	Student progress towards the game goal, the rate and change of this progress
[15]	Fo, Ca	Dv	Ad, Pz	Sg	Yes	App	Students' responses
[30]	It, Ca	Dv	Ad	Sg	Yes	N/A	Student progress in the game, score
[31]	Un	Dm	Si	Sg	Yes	N/A	Game completion results, game performance score
[32]	Un	Dm	St	Sg	No	N/A	Score, Used game characters, Time viewing information
[33]	Un	Dm, Dv	Rp	Sg	Yes	App	Interacting with game tools/elements, time, score, Taken game path
[34]	Un	Dm	Rp	Sg	Yes	N/A	Interacting with game tools/elements, path follow, time reading story
[35]	Un	Dm	Rp	Sg	No	N/A	Agents' pedagogical discourse moves, Cognitive-discourse variables of the student
[36]	Un	Dm, Lg	Rp	Mm	No	N/A	Teaming up, engaging in battles, learning, trading, interacting with game elements or tools, and chatting
[37]	Un	Dm	Ad	Sg	No	N/A	Game performance score
[38]	Un	Dm	Pz	Sg	No	N/A	Students' given solutions and attempts
[39]	Un	Dm	Pz	Sg	Yes	N/A	Number of clicks, Discovery /research duration
[40]	Un	Dv	Rp	Sg	No	App	Number of accessed game tools
[41]	Un	Dm	Pz	Sg	No	N/A	Student's personal information, Number of correct and wrong answers, Score, and number of gestures
[42]	Un	Dm	Pz	Sg	No	N/A	Created fractions
[43]	Un	Lg	Si	Sg	No	N/A	Note-taking, Interaction with game tools, Game time
[44]	Fo	Dv	N/A	N/A	Yes	N/A	N/A
[45]	Fo, It	Dv	Rp	Sg	Yes	N/A	Time spent in each game scene, Game location, Reached game level
[46]	Fo, It	Dv	Si	Sg	Yes	N/A	Game interaction time, interacting with game tools, game time
[47]	Fo	Dm	Pz	Sg	No	N/A	Students' given solutions and attempts
[48]	Fo	Dm	Pz	Sg	No	App	Game score, Fight Game score
[49]	Fo	Dm, Dv	N/A	N/A	Yes	N/A	N/A
[50]	It	Dm	FPS	Sg	No	N/A	Game time and levels, score, number of deaths
[51]	It	Dv	Rp	Sg	Yes	N/A	Students' responses and game locations
[52]	It, Ca	Dm	Rp	Sg	Yes	N/A	Students' responses
[53]	It, Ip	Dm	Rp	Sg	Yes	N/A	Students' responses
[54]	It	Dv	Pz	Sg	Yes	App	Number of mistakes, Time solving a level
[55]	Ca	Dv	Si	Mp	Yes	N/A	Game score, time answering quests
[56]	Ca	Dm	Si	Mp	No	N/A	Chat and Forum traces
[57]	Ca	Dm	Pz	Sg	No	N/A	Number of tries solving the quest, Hand movement
[58]	Ca	Dm	Rp	Sg	Yes	N/A	Coordinates of movement, time-stamps, special
[59]	Ca	Dm	Rp	Sg	Yes	N/A	Game events, number of villagers retrieved
[60]	Ca	Dm	Rp	Sg	Yes	N/A	Coordinates of movement, time-stamps, special game events, number of villagers retrieved
[61]	Ca	Dm, Dv	Pz	Sg	Yes	App	Students' errors
[62]	Ca	Dm, Dv	Si	Sg	No	N/A	Student name and age, Correct/wrong answer, game time
[63]	Ca	Dm	Pz	Sg	No	N/A	Interaction with game tools/items

[a] LA Goal (Un = Understanding and modeling students' in-game behaviors, Fo = Formative design of educational games, It = Implementing teaching support, Ca = Conducting learning assessment, Ip = In-game personalization), [b] LA approach (Dm = Data mining, Lg = Lag sequential analysis, Dv = Data visualization), [c] Game Genre (Puzzle = Pz, Adventure = Ad, Roleplaying = Rp, Strategy = St, Simulation = Si, First person shooting = FPS, Not applicable = N/A), [d] Game Mode (Single player = Sg, Multiplayer = Mp, Massively Multiplayer, = Mm), [e] Embedded Analytics (Yes or No), [f] Game Traces, [g] Challenges (Applicable = App, Non applicable = N/A).

References

[1] W.-C. Hsu and H.-C. K. Lin, “Impact of applying WebGL technology to develop a web digital game-based learning system for computer programming course in flipped classroom,” in *International Conference on Educational Innovation through Technology (EITT)*, Tainan, Taiwan, China, 2016, doi: 10.1109/EITT.2016.20.

[2] F. Ke., “An implementation of design-based learning through creating educational computer games: A case study on mathematics learning during design and computing,” *Computers & Education*, vol. 73, pp. 26-39, 2014, doi: 10.1016/j.compedu.2013.12.010.

[3] T.-Y. Liu and Y.-L. Chu, “Using ubiquitous games in an English listening and speaking course: Impact on learning outcomes and motivation,” *Computers & Education*, vol. 55, no. 2, pp. 630-643, 2010, doi: 10.1016/j.compedu.2010.02.023.

[4] S. Suh, S. W. Kim, and N. J. Kim, “Effectiveness of MMORPG-based instruction in elementary English education in Korea,” *Journal of Computer Assisted Learning*, vol. 26, no. 5, pp. 370-378, 2010, doi: 10.1111/j.1365-2729.2010.00353.x.

[5] V. J. Shute, L. Wang, S. Greiff, W. Zhao, and G. Moore, “Measuring problem solving skills via stealth assessment in an engaging video game,” *Computers in Human Behavior*, vol. 63, pp. 106-117, 2016, doi: 10.1016/j.chb.2016.05.047.

[6] D. Burgos, C. Tattersall, and R. Koper, “Re-purposing existing generic games and simulations for e-learning,” *Computers in Human Behavior*, vol. 23, no. 6, pp. 2656-2667, 2006, doi: 10.1016/j.chb.2006.08.002.

[7] P. Moreno-Ger and D. Burgos, “The case for serious games analytics,” in *Radical Solutions and Learning Analytics: Personalised Learning and Teaching Through Big Data*, D. Burgos (Ed.), Singapore: Springer, 2020, pp. 213-227, doi:10.1007/978-981-15-4526-9_13.

[8] D. Burgos, *Radical solutions and learning Analytics*, Singapore: Springer, 2020.

[9] H. Fournier, R. Kop, and G. Durand, “Challenges to research in MOOCs,” *Journal of Online Learning and Teaching / MERLOT*, vol. 10, no. 1, pp. 1–15, 2014.

[10] J. Frommel, K. Rogers, J. Brich, D. Besserer, L. Bradatsch, I. Ortinau, ... and M. Weber, Integrated questionnaires: Maintaining presence in game environments for self-reported data acquisition, in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, London, United Kingdom, 2015, pp. 359-368.

[11] D. D. Reese, B.G. Tabachnick, and R. E. Kosko, “Video game learning dynamics: Actionable measures of multidimensional learning trajectories,” *British Journal of Educational Technology*, vol. 46, no. 1, pp. 98-122, 2015, doi: 10.1111/bjet.12128

[12] V. J. Shute and M. Ventura, *Stealth assessment: Measuring and supporting learning in video games*, London, England: The MIT Press, 2013.

[13] B. R. Belland, “The role of construct definition in the creation of formative assessments in game-based learning,” in *Assessment in game-based learning*, New York, United States: Springer, 2012, pp. 29-42, doi:10.1007/978-1-4614-3546-4_3

[14] V. J. Shute, M. Ventura, M. Bauer, and D. Zapata-Rivera, “Melding the power of serious games and embedded assessment to monitor and foster learning,” in *Serious games: Mechanisms and effects*, New York, United States: Routledge, 2009, pp. 295-321.

[15] Á. Serrano-Laguna, J. Torrente, P. Moreno-Ger, and B. Fernández-Manjón. “Application of LA in educational videogames,” *Entertainment Computing*, vol. 5, no. 4, pp. 313-322. 2014. doi:10.1016/j.entcom.2014.02.003

[16] I. J. Perez-Colado, V. M. Perez-Colado, I. Martínez-Ortiz, M. Freire-Moran, and B. Fernández-Manjón, “UAdventure: The eAdventure reboot: Combining the experience of commercial gaming tools and tailored educational tools,” in *IEEE Global Engineering Education Conference (EDUCON)*, Athens, Greece, 2017.

[17] A. Tlili, F. Essalmi, M. Jemni, Kinshuk, and N.S. Chen, “A complete validated learning analytics framework: Designing issues from data preparation perspective,” *International Journal of Information and Communication Technology Education (IJICTE)*, vol. 14, no. 2, pp. 1–16, 2018, doi: 10.4018/IJICTE.2019070104.

[18] H. Fournier, R. Kop, and H. Sitlia, “The value of learning analytics to networked learning on a personal learning environment,” in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge,* Banff, Canada, 2011, pp. 104-109, doi:10.1145/2090116.2090131.

[19] S. Powell and S. MacNeill, “Institutional readiness for analytics,” *JISC CETIS Analytics Series*, vol. 1, no. 8, 2012.

[20] D. Gijbels and F. Dochy, “Students’ assessment preferences and approaches to learning: can formative assessment make a difference?,” *Educational Studies*, vol. 32, no. 4, pp. 399-409. 2006. doi: 10.1080/03055690600850354.

[21] V. J. Shute and M. Ventura, “Stealth assessment,” in *The SAGE Encyclopedia of Educational Technology*, Thousand Oaks, United States: SAGE, 2015, pp. 675-678, doi: 10.4135/9781483346397.

[22] C.C. van Nimwegen, D. Burgos, H.H. van Oostendorp and H.H. Schijf, “The paradox of the assisted user: guidance can be counterproductive” in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, Montreal, Canada, 2006, pp.917-926.

[23] C. Alonso-Fernandez, A. Calvo, M. Freire, I. Martinez-Ortiz, and B. Fernandez-Manjon. “Systematizing game learning analytics for serious games” *in IEEE Global Engineering Education Conference (EDUCON),* Athens, Greece, 2017, pp. 1111-1118.

[24] A. Tlili and M. Chang, “Data analytics approaches in educational games and gamification Systems: Summary, challenges, and future insights,” in *Data Analytics Approaches in Educational Games and Gamification Systems*, Singapore: Springer, 2019, pp. 249-255, doi:10.1007/978-981-32-9335-9

[25] Z. K. Papamitsiou and A.A. Economides, “Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence,” *Journal of Educational Technology & Society*, vol. 17, pp. 49-64, 2014.

[26] G. L. Saveski, W. Westera, L. Yuan, P. Hollins, B.F. Manjón, P.M. Ger, and K. Stefanov, “What serious game studios want from ICT research: identifying developers’ needs,” in *International Conference on Games and Learning Alliance*, Rome, Italy, 2015, pp. 32-41.

[27] C. Okoli and K. Schabram, “A guide to conducting a systematic literature review of information systems research,” *Sprouts: Working Papers on Information Systems*, vol. 10, no. 26, 2010, doi: 10.2139/ssrn.1954824.

[28] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, “A systematic literature review of empirical evidence on computer games and serious games,” *Computers & Education*, vol. 59, no. 2, pp. 661-686, 2012, doi: 10.1016/j.compedu.2012.03.004.

[29] J. Webster and R. T. Watson, “Analyzing the past to prepare for the future: Writing a literature review,” *MIS Quarterly*, vol. 26, no. 2, pp. 13-23, 2002.

[30] M. Minović, M. Milovanović, U. Šošević, and M. Á. C. González. “Visualisation of student learning model in serious games,” *Computers in Human Behavior*, vol. 47, pp. 98-107, 2015, doi:10.1016/j.chb.2014.09.005.

[31] C. Alonso-Fernández, I. Martínez-Ortiz, R. Caballero, M. Freire, and B. Fernández-Manjón, “Predicting students’ knowledge after playing a serious game based on learning analytics data: A case study,” *Journal of Computer Assisted Learning*, vol. 36, no. 3, 2019, doi: 10.1111/jcal.12405.

[32] M. T. Cheng, Y. W. Lin and H. C. She, “Learning through playing virtual Age: Exploring the interactions among student concept learning, gaming performance, in-game behaviors, and the use of in-game characters,” *Computers & Education*, vol. 86, pp. 18-29, 2015, doi:10.1016/j.compedu.2015.03.007

[33] M. Denden, A. Tlili, F. Essalmi, and M. Jemni, “Implicit modeling of learners’ personalities in a game-based learning environment using their gaming behaviors,” *Smart Learning Environments*, vol. 5, no. 1, 2018, doi:10.1186/s40561-018-0078-6

[34] F. Essalmi, A. Tlili, L. J. B. Ayed, and M. Jemni. “Toward modeling the learner’s personality using educational games,” *International Journal of Distance Education Technologies (IJDET),* vol. 15, no. 4, pp. 21-38, 2017.

[35] C. M. Forsyth, A. C. Graesser, P. Pavlik Jr, Z. Cai, H. Butler, D. Halpern, and K. Millis. “Operation aries!: Methods, mystery, and mixed models: Discourse features predict affect in a serious game,” *Journal of Educational Data Mining*, vol. 5, no. 1, pp. 147-189, 2013, 10.5281/zenodo.3554615

[36] H. T. Hou, “Exploring the behavioral patterns of learners in an educational massively multiple online role-playing game (MMORPG),” *Computers & Education*, vol. 58, no. 4, pp. 1225-1233, 2012, doi: 10.1016/j.compedu.2011.11.015.

[37] R. Israel-Fishelson and A. Hershkovitz, “Persistence in a game-based learning environment: The case of elementary school students learning computational thinking,” *Journal of Educational Computing Research*, vol. 58, no. 5, pp. 891-918, 2019, doi: 10.1177/0735633119887187.

[38] D. Kerr and G. K. Chung. "Identifying key features of student performance in educational video games and simulations through cluster analysis," *Journal of Educational Data Mining*, vol. 4, no. 1, pp. 144-182, 2012, doi:10.5281/zenodo.3554647

[39] M. A. Khenissi, F. Essalmi, M. Jemni, T. W. Chang, and N. S. Chen. "Unobtrusive monitoring of learners' interactions with educational games for measuring their working memory capacity," *British Journal of Educational Technology*, vol. 48, no. 2, pp. 224-245, 2017, doi:10.1111/bjet.12445

[40] M. Liu, J. Lee, J. Kang, and S. Liu. "What we can learn from the data: A multiple-case study examining behavior patterns by students with different characteristics in using a serious game," *Technology, Knowledge and Learning*, vol. 21, no. 1, pp. 33-57, 2016, doi: 10.1007/s10758-015-9263-7.

[41] E. Lotfi, B. Amine, and B. Mohammed. "Players performances analysis based on educational data mining case of study: Interactive waste sorting serious game," *International Journal of Computer Applications*, vol. 108, no. 11, pp. 13-19, 2014.

[42] T. Martin, C. Petrick Smith, N. Forsgren, A. Aghababyan, P. Janisiewicz, and S. Baker. "Learning fractions by splitting: Using learning analytics to illuminate the development of mathematical understanding," *Journal of the Learning Sciences*, vol. 24, no. 4, pp. 593-637, 2015, doi: 0.1080/10508406.2015.1078244.

[43] C. T. Wen, C. J. Chang, M. H. Chang, S. H. F. Chiang, C. C. Liu, F. K. Hwang, and C. C. Tsai. "The learning analytics of model-based learning facilitated by a problem-solving simulation game," *Instructional Science*, vol. 46, no. 6, pp. 847-867, 2018, doi:10.1007/s11251-018-9461-5.

[44] G. Altanis, M. Boloudakis, S. Retalis, and N. Nikou. "Children with motor impairments play a kinect learning game: First findings from a pilot case in an authentic classroom environment," *Journal of Interact Design Architect*, vol. 19, pp. 91-104, 2013.

[45] A. Calvo-Morata, D. C. Rotaru, C. Alonso-Fernández, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón. "Validation of a cyberbullying serious game using game analytics," *IEEE Transactions on Learning Technologies*. vol. 13, no. 1, pp. 186-197, 2018.

[46] A. R. Cano, B. Fernández-Manjón, and Á. J. García-Tejedor. "Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities," *British Journal of Educational Technology*, vol. 49, no. 4, pp. 659-672, 2018.

[47] D. Kerr. "Using data mining results to improve educational video game design," *Journal of Educational Data Mining*, vol. 7, no. 3, pp. 1-17, 2015.

[48] Á. Serrano-Laguna, B. Manero, M. Freire, and B. Fernández-Manjón. "A methodology for assessing the effectiveness of serious games and for inferring player learning outcomes," *Multimedia Tools and Applications*, vol. 77, no. 2, pp. 2849-2871, 2018. doi:10.1007/s11042-017-4467-6

[49] Y. Chaudy and T. Connolly. "Specification and evaluation of an assessment engine for educational games: Integrating learning analytics and providing an assessment authoring tool," *Entertainment Computing*, vol. 30, 2019, doi: 10.31686/IJIER.VOL8.ISS3.2220.

[50] M. Callaghan, N, McShane, A. G. Eguíluz, and M. Savin-Baden. "Extending the activity theory based model for serious games design in engineering to integrate analytics," *International Journal of Engineering Pedagogy (iJEP),* vol. 8, no. 1, pp. 109-126, 10.3991/ijep.v8i1.8087.

[51] Z. H. Chen and S. Y. Lee. "Application-driven educational game to assist young children in learning english vocabulary," *Journal of Educational Technology & Society*, vol. 21, no. 1, pp. 70-81, 2018.

[52] K. Kiili and H. Ketamo. "Evaluating cognitive and affective outcomes of a digital game-based math test," *IEEE Transactions on Learning Technologies*, vol. 11, no. 2, pp. 255-263, 2018.

[53] K. Kiili, K. Moeller, and M. Ninaus, "Evaluating the effectiveness of a game-based rational number training-In-game metrics as learning indicators," *Computers & Education*, vol. 120, pp. 13-28, 2018, doi: 10.1016/j.compedu.2018.01.012.

[54] D. Rodríguez-Cerezo, A. Sarasa-Cabezuelo, M. Gómez-Albarrán, and J. L. Sierra, "Serious games in tertiary education: A case study concerning the comprehension of basic concepts in computer language implementation courses," *Computers in Human Behavior*, vol. 31, pp. 558-570, 2014, doi: 10.1016/j.chb.2013.06.009.

[55] A. Capatina, G. Bleoju, E. Rancati, and E. Hoareau, "Tracking precursors of learning analytics over serious game team performance ranking," *Behaviour & Information Technology*, vol. 37, no. 10-11, pp. 1008-1020, 2018, doi:10.1080/0144929X.2018.1474949.

[56] A. B. Hernández-Lara, A. Perera-Lluna, and E. Serradell-López. "Applying learning analytics to students' interaction in business simulation games: The usefulness of learning analytics to know what students really learn," *Computers in Human Behavior*. vol. 92, pp. 600-612. 2018, doi: 10.1016/j.chb.2018.03.001.

[57] M. Kourakli, I. Altanis, S. Retalis, M. Boloudakis, D. Zbainos, and K. Antonopoulou. "Towards the improvement of the cognitive, motoric, and academic skills of students with special educational needs using Kinect learning games," *International Journal of Child-Computer Interaction*, vol. 11, pp. 28-39, 2017, doi:10.1016/j.ijcci.2016.10.009.

[58] C. S. Loh and Y. Sheng. "Maximum similarity index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games," *Computers in Human Behavior*, vol. 39, pp. 322-330, 2014, doi: 10.1016/j.chb.2014.07.022.

[59] C. S. Loh and Y. Sheng. "Measuring the (dis-) similarity between expert and novice behaviors as serious games analytics," *Education and Information Technologies*, vol. 20, no. 1, pp. 5-19, 2015, doi: 10.1007/s10639-013-9263-y.

[60] C. S. Loh, Y. Sheng, and I. H. Li. "Predicting expert-novice performance as serious games analytics with objective-oriented and navigational action sequences," *Computers in Human Behavior*, vol. 49, pp. 147-155, 2015. doi: 10.1016/j.chb.2015.02.053.

[61] E. Rowe, J. Asbell-Clarke, R. S. Baker, M. Eagle, A. G. Hicks, T. M. Barnes, and T. Edwards. "Assessing implicit science learning in digital games," *Computers in Human Behavior*, vol. 76, pp. 617-630, 2017, doi: 10.1016/j.chb.2017.03.043.

[62] A. Slimani, F. Elouaai, L. Elaachak, O. B. Yedri, M. Bouhorma, and M. Sbert. "Learning analytics through serious games: Data mining algorithms for performance measurement and improvement purpose," *International Journal of Emerging Technologies in Learning (iJET),* vol. 13, no. 1, pp. 46-64, 2018.

[63] E. L. Snow, L. K. Allen, M. E. Jacovina, and D. S. McNamara. "Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment," *Computers & Education*, vol. 82, pp. 378-392, 2015. doi: 10.1016/j.compedu.2014.12.011.

[64] K. Becker, "How are games educational? Learning theories embodied in games," in *DiGRA: Changing Views - Worlds in Play*. Vancouver, Canada, 2005.

[65] E. Z. F., Liu, and C. H. Lin. "Developing evaluative indicators for educational computer games," *British Journal of Educational Technology*, vol. 40, no. 1, pp. 174-178, 2009, doi:10.1111/j.1467-8535.2008.00852.x.

[66] J. Moon and Z. Liu. "Rich Representations for Analyzing Learning Trajectories: Systematic Review on Sequential Data Analytics in Game-Based Learning Research," in *Data Analytics Approaches in Educational Games and Gamification Systems*, Tlili A., Chang M. (eds), Singapore: Springer, 2019.

[67] A. Tlili, F. Essalmi, and M. Jemni, "An educational game for teaching computer architecture: Evaluation using learing analytics," in *5th IEEE International Conference on Information & Communication Technology and Accessibility (ICTA)*, Marrakech, Morocco, 2016, pp. 1-6.

[68] A. Pardo and G. Siemens, "Ethical and privacy principles for learning analytics," *British Journal of Educational Technology*, vol. 45, no. 3, pp. 438–450, 2014, doi: 10.1111/bjet.12152.

[69] S. Welsh and S. McKinney, "Clearing the Fog: A learning Analytics Code of Practice," in *Proceedings of the Australasian Society for Computers in Learning in Tertiary Education*, T. Reiners, B.R. von Konsky, D. Gibson, V. Chang, L. Irving, & K. Clarke (Eds.), Perth, Australia. 2015.

[70] S. Erhel and E. Jamet. "Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness," *Computers & Education*, vol. 67, pp. 156-167, 2013, doi: 10.1016/j.compedu.2013.02.019.

Ahmed Tlili

Ahmed Tlili is the Co-Director of the OER Lab at the Smart Learning Institute of Beijing Normal University (SLIBNU), China. He serves as the Associate Editor of the IEEE Bulletin of the Technical Committee on Learning Technology, and the Journal of e-Learning and Knowledge Society. He is also a Visiting Professor at UNIR-iTED, Spain, and an expert at the Arab League Educational, Cultural and Scientific Organization (ALECSO). Dr. Tlili has been awarded the IEEE TCLT Early Career Researcher Award in Learning Technologies for 2020. He has edited several special issues in indexed journals. He has also published several books, as well as academic papers in international referred journals and conferences. He is serving as a local organizing and program committee member in various international conferences, and as a reviewer in several refereed journals. Dr. Tlili is the Co-Chair of IEEE special interest group on "Artificial Intelligence and Smart Learning Environments" and APSCE's Special Interest Group on "Educational Gamification and Game-based Learning (EGG)". His research interests include, open education, game-based learning, educational psychology and artificial intelligence.

Maiga Chang

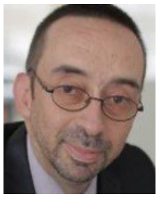
Maiga Chang is currently a Full Professor with the School of Computing Information and Systems, Athabasca University, Canada. He has given more than 105 talks and lectures in different conferences, universities, and events. He has participated in more than 310 international conferences and workshops as a Program Committee Member. He has (co-)authored more than 225 edited books, special issues, book chapters, journal and international conference papers. He is the Editor-in-Chief of the Educational Technology and Society, the Bulletin of Technical Committee on Learning Technology, and the International Journal of Distance Education Technologies. He was a Section Editor of the Education and Science, an Associate Editor of Transactions on Edutainment (Springer). He is an Advisory Board Member of the Journal of Computers and Applied Science Education. He is also the Chair of the IEEE Technical Committee of Learning Technology (IEEE TCLT), an Executive Committee Member of the Asia-Pacific Society for Computers in Education (APSCE), the Global Chinese Society for Computing in Education (GCSCE), and the Chinese Society for Inquiry Learning (CSIL). He is a Secretary and a Treasurer of the International Association of Smart Learning Environments (IASLE).

Jewoong Moon

Jewoong Moon is a Ph.D. candidate at the Department of Educational Psychology and Learning Systems at Florida State University, United States. His research interests include digital game-based learning, immersive learning environment design for neurodiverse learners, learning analytics & educational data mining, and adaptive learning system design. He is currently investigating multimodal data fusion implementation to develop predictive models as stealth assessment for adaptive and immersive learning environments.

Zhichun "Lukas" Liu

Dr. Zhichun "Lukas" Liu is a postdoctoral fellow at Kaput Center of Research and Innovation in STEM Education, University of Massachusetts Dartmouth. His current work aims at using game-based learning to promote the development and learning transfer of computational thinking among K-12 students and teachers. His research interests also include using quantitative analytics methods to understand learners' behavior, competence, discourse, and interactions.

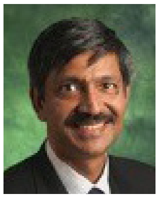


Daniel Burgos

Daniel Burgos received a postgraduate in artificial intelligence & machine learning from MIT, and the Ph.D. degree in communication, the Dr.Ing. degree in computer science, the Ph.D. degree in education, the Ph.D. degree in anthropology, and the D.B.A. degree in business administration. He works currently as a Full Professor of technologies for education & communication and Vice-Rector for International Research, UNESCO Chair of eLearning, and ICDE Chair of open educational resources at Universidad Internacional de La Rioja. He is also the Director of the Research Institute for Innovation & Technology in Education (UNIR iTED). Further, he is a Professor with An-Najah National University, Palestine, an Adjunct Professor with the Universidad Nacional de Colombia (UNAL), Colombia, an Extraordinary Professor with North-West University, South Africa, a Visiting Professor with Coventry University, U.K., and the Universidad de las Fuerzas Armadas (ESPE), Ecuador. He has published over 150 scientific articles, 20 books, and 15 special issues on indexed journals. He has developed +55 European and Worldwide Research and Development projects. His research interests include adaptive, personalised and informal eLearning, learning analytics, open education and open science, eGames, and eLearning specifications.

Nian-Shing Chen

Nian-Shing Chen is currently Chair Professor in the Department of Applied Foreign Languages at the National Yunlin University of Science and Technology, Taiwan. He has published over 400 academic papers in the international referred journals, conferences and book chapters. One of his papers published in Innovations in Education and Teaching International was awarded as the top cited article in 2010. He is the author of three books with one textbook entitled "e-Learning Theory & Practice". He has received the national outstanding research awards for three times from the National Science Council in 2008, 2011-2013 and the Ministry of Science and Technology in 2015-2017. His current research interests include assessing e-Learning course performance; online synchronous teaching & learning; mobile & ubiquitous learning; gesture-based learning and educational robotics.



Kinshuk

Kinshuk is currently a Professor of Computer Science and the Dean of the College of Information at the University of North Texas, USA. He received the Ph.D degree in computer science from the University of De Montfort, England, in 1996. He held the NSERC/CNRL/Xerox/McGraw Hill Research Chair for Adaptivity and Personalization in Informatics, funded by the Federal government of Canada, Provincial government of Alberta, and by national and international industries. Area of his research interests include learning analytics; learning technologies; mobile, ubiquitous and location aware learning systems; cognitive profiling; and interactive technologies.

Local Technology to Enhance Data Privacy and Security in Educational Technology

Daniel Amo[1]*, Paul Prinsloo[2], Marc Alier[3], David Fonseca[1], Ricardo Torres Kompen[1], Xavier Canaleta[1], Javier Herrero-Martín[4]

[1] La Salle, Ramon Llull University, Barcelona (Spain)
[2] University of South Africa, Pretoria (South Africa)
[3] Polytechnical University of Catalonia, Barcelona (Spain)
[4] La Salle University Center, Autonomous University of Madrid, Madrid (Spain)

Received 26 February 2021 | Accepted 22 April 2021 | Published 23 November 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

In educational environments, technological adoption in the last 10 years has enabled a data-driven and decision-making paradigm in organizations. The integration of cloud services in schools and universities is a positive shift in the field of learning, but it also presents threats to all academic roles that need to be discussed in terms of protection, privacy, and confidentiality. Cloud storage brings the ubiquity of data to this technical transition and a delusive opportunity for cost savings. In many cases, this suggests that certain actors, beyond the control of schools and colleges, collect, handle and treat educational data on private servers and data centers. This privatization enables the manipulation of stored records, leaks, and unauthorized access. In this article, we expose the possibilities that open from the viewpoint of local technology adoption. We seek to reduce or even totally solve the detrimental effects of using cloud-based instructional and analytical technology, mixing or only using local technology. Technological methods that conform to this alternate viewpoint and new lines of study are also being suggested and created.

Keywords

Cloud Computing, Education, Local Technologies, Privacy, Security.

DOI: 10.9781/ijimai.2021.11.006

I. Introduction

Schools and universities are experiencing rapid processes of digital updating where the adoption of third-party technology solutions results in changes in academic and learning processes. This digital evolution means the use of third-party hardware and software, which mostly resides and executes in cloud computing [1], [2], and forces changes to educational institutions. On the one hand, institutions require permanent Internet connectivity, and therefore the data generated becomes ubiquitous and available at any time and anywhere. On the other hand, changes in academic and educational processes raise new problems concerning generated data. Consequently, the new digital learning context is causing a profound change in educational organizations as well as new issues that need to be addressed.

The growth of educational technology based on cloud computing has led to the adoption of educational decisions based on data in line with the big data analytics movement [3]. This type of hyperconnected educational environment has a strong ability to collect, store, process, and analyze large amounts of data through cloud computing. Technological innovations and the cheapening of cloud computing have made Software as a Service (SaaS) [4], which resides in the cloud, the most attractive option for the distribution of digital tools in the industrial field. Even in educational terms, the philosophy of cloud computing and services also applies [5].

Business models seek IT solutions to save costs. Cloud computing helps to maintain an architectural design geared towards storing and processing large volumes of personal data, data, and metadata on remote servers. The educational context does not escape the adoption of cloud computing; universities and schools are working to reduce costs [6]. However, the use of the cloud in this context results in many negative results, from data leaks to misuse. Consequently, digital educational technology tools (EdTech) raise new challenges and issues related to privacy, identity, confidentiality, and security of data and metadata (PICSDM) for all educational roles and actors involved [7]-[13]. We need to ask ourselves:

- Do we need to send data to cloud computing?
- What data should be sent outside the institution and can be transferred without risking the PICSDM?
- Do we need to send unsecured data to cloud computing, or can we send it anonymously?
- Can local technology provide a trustworthy, private, and secure environment?

If it is necessary and justified to send data to cloud computing, it is mandatory to define and integrate both processes and technology that ensure ethical and legal treatment.

* Corresponding author.
E-mail address: daniel.amo@salle.url.edu

We present a framework named LEDA (Local Educational Data Analytics) and fathom into the "local first" principle. The LEDA framework builds in of seven principles with a big focus on considering local technology in any analytical solution design. Thus, in this work we make an extensive exposition of the LEDA framework to show how local technology should be considered, studied, and debated as one more to solve in new ways, or minimize, privacy problems known in analytical educational environments.

The LEDA framework principles enhance data privacy and security in the development or adoption of digital tools in education. First, it encourages the use of the "local technology without data transfer outside the classroom" premise, not excluding the "remote with data transfer to cloud computing" premise but relegating it to the last place. Between these two premises is a very wide range of options. Technologically you can work locally, integrate a mixed system that stores personal data locally and links to systems in cloud computing, to work absolutely in cloud computing. The authors of this paper seem to be against cloud computing. On the contrary, our proposal lies in inferring from those ultimately responsible for technological infrastructures a new vision of new perspectives and possibilities concerning the local-cloud binomial:

- First, indicate that the concept of "local" does not refer only to servers of the institution. Both the institutional services and the educational staff can work with technological solutions that can reside both in the servers of the institution and in the devices provided to teachers.
- Second, we propose that the premises be considered within the technological equation as one more option. Every educational institution has limited resources and it may seem a natural evolution to solve any short-term problem via the adoption of cloud computing. However, working locally can activate a distributed architecture within the institution that achieves the same results as a centralized architecture in cloud computing.
- Third, the concept of "local" is not opposed to cloud computing. Only local, mixed solutions between local and cloud computing, or only cloud computing can coexist in an institutional space. We want to prevent cloud computing from being used as a unique solution where the keys are the problems to be solved.

As mentioned, the "local first" principle is the most important of the LEDA framework. In the following sections, we will justify in what ways it contributes something different by opening new possibilities of doing not considered until now. We understand the rest of the principles as the necessary context that supports it from a legal and ethical perspective. That is why in this work we focus on exposing the "local first" principle extensively to make the benefits visible and encouraging people to turn their attention to local technology.

The structure of the paper consists of six sections. The first section is the introduction. In the second section, we present the pedagogical and technological transformation of the La Salle institution as the trigger of the LEDA principles development, and the risks raised with technology innovation in cloud computing as well. In the third section, we present the seven principles that underpin the framework. In the fourth section, we present how the local principle can solve in new ways some problems, expose some technical possibilities to develop "local first" principle compliance technologies, and show some actual solutions already developed or in development as well. In the next fifth section, we present both future and ongoing lines of research. Finally, we summarize the conclusions of the work as the last and sixth current section.

II. Context

A. Act Locally, Think Globally

The ARLEP district (Agrupación Lasaliana España - Portugal) of La Salle institution in 2017 began the design of a new pedagogical framework called NCA [14], [15] and in 2019 started the deployment of the framework among all their schools and universities. This pedagogical transformation provides new methodological, new didactic formulas, and an intense innovation through the use of digital technology. Related to the uses of digital technology in cloud computing some data risks emerge. In this risky context, the LEDA framework sets the principles to ensure privacy, identity, confidentiality, and security of your data, personal data, and metadata (PICSDM) in the NCA transformation acting locally but think globally considering solutions that can be beneficial to any educational institution of the world and beyond.

At the same time, this framework initiates different research related to digital competence and data literacy. Within digital competence, we find data literacy as the ability to interpret and analyze data. In educational settings, this knowledge is useful for both teachers and students to make data-driven decisions considering that it is an increasingly technicalization environment and surrounded by data. The research will extract indicators of the state of the situation on digital knowledge and data literacy as well as the creation of self-assessment tools and specific training. It is hoped that these indicators, self-assessment tools, and training will raise awareness of the various issues related to the sensitivity of data in educational settings and active action to prevent them.

B. Data as Sensible Educational Assets

The collection, storage, and treatment of educational data by third actors cause undesired situations that make them very sensitive and fragile. This context is due to the uncontrolled introduction of Big Data technology, consisting of Artificial Intelligence [16], Machine Learning, Deep Learning, and Neural Networking techniques in combination with Cloud Computing. This set of technologies and techniques applied to education has indeed revolutionized, especially in decision automation. However, it has also led to a strong distrust of educational institutions, as they correspond to a context with many open questions and loss of control. Some examples of these analytical technologies are the massive automatic decision-making, massive sensible data collection from students [12], unauthorized access to data [17], discrimination, filtering, analysis, and predictive tools against students will [18], and data transfer without a legally defined relationship [9]. There is an unstable situation in collecting, treating, sharing, and analyzing educational data [13].

Moreover, Cloud Computing enables Dark Data. Dark Data is all the collected but unused data. This data is stored waiting to be used for the benefit of technology. At first, it is not known how they will be used. That is why Dark Data presents an uncertain and probably frightening future if current issues are considered.

Data is fragile whether in the Cloud or anywhere else. Data may be more fragile in the Cloud because there are more actors involved. However, locally stored data is also fragile and open for misinterpretation and misuse. Considering this provides a different perspective on the collection, storage, and treatment of data, while it raises the importance of ethical behaviors. "Author et al." [19] delve into the concept of sensitive data and data fragility. They pointed out that there is definitely (more) sensitive data. Depending on the context, any data can become sensitive, e.g., location data, devices version, or any previous academic performance can depend on the context, escalate into sensitivity. What is explicating "Author et al." is metadata. This is the reason why the authors will always specify data as both personal data and metadata, considering metadata all interactions and kind of other data than personal associated with them.

1. Power Asymmetries

It is a rapidly evolving trend to develop digital educational tools in cloud computing. The problem with these tools, such as ClassDojo or Snappet, as different authors point out [7], [11], [20], is that they generate situations of active surveillance and even manipulation. It means that they somehow modify the environment to generate arbitrary behaviors in students, which self-regulates the power-technology-teaching-learning relationships in an unbalanced way. As Williamson [21] points out, these technologies are governed by automatic decision-making algorithms [22], developed with Big Data technologies, finally harmful by regulating educational processes thanks to the ability to execute predictive algorithms in cloud computing.

Two examples of this are the analysis by Norwegian Consumer Councils and the Norwegian Data Inspectorate. On the one hand, Norwegian Consumer Councils highlights the Dark Patterns those technologies such as Google, Facebook, or Microsoft use to reduce the privacy options on their devices [23], and even forcing the users to accept being tracked continuously [24]. These situations generate an asymmetry of powers that tilts the balance in favor of the technology companies 'profits, leaving users who use their services at their mercy and without many desirable privacy settings. In general, there are no devices intended exclusively for education. Students use generic devices such as personal computers or smartphones from companies that apply dark patterns. Therefore, students are victims of dark patterns without being able to opt for devices that protect them from these asymmetrical and monopolistic practices. On the other hand, the guide published by the Norwegian Data Inspectorate is another example in an educational context of how to monitor and profile students when using Google tools [25]. Google has a specific license for education that specifies which tools and services the GDPR complies with. This license protects students and prevents them from being profiled to serve advertising on certain tools. ChromeOS, Google's operating system (Chromebooks), is not included in this license. However, Google promotes Chromebooks as suitable devices for education when Norwegian Data Inspectorate points to the opposite.

Algorithms are not neutral and are filled with cognitive biases [26] weakening the confidence in their adoption, which in the case of education is based on Learning Analytics [26]–[28], which offer visual data analytics to reduce data literacy and enhance educational processes, such as evaluation [30]. This is why the use of Learning Analytics has caused concern and generated debate [31], [32], giving way to the proposal of frameworks, guidelines, and recommendations for use such as DELICATE [11] and SHEILA [33]. Despite the great effort being made to restore confidence, regulate, and make the integration of Learning Analytics into EdTech ethical and ensure PICSDM, it can be shown that cases that degrade confidence in the use of algorithms and analysis of educational data continue to happen. The case of the A-levels in the UK [34] is an example where algorithms applied by the government have undermined the qualification of a large majority of students by preventing them from accessing the desired university studies. These disastrous results in the use of algorithms in education have aroused great disagreement among students and a strong rejection of such practices [34]; millions of pieces of sensitive student data have been leaked in recent years [35] relating to personal data, registration data and even financial data [36].

By now, it is clear that a "human in the loop" is needed [37]. Moore contrasts the concept of "algorithmic decision" with "algorithmic results" when she says that "… algorithmic decision not least that there is no human outside to act as the guarantor of the good…" but that "… they're always already also inside that new framework or paradigm of knowledge so then there's no decision as such in what societies have begun to call algorithmic decision, there are outputs…". Dark patterns, no educational-oriented devices, no human-in-the-loop, and active decisional biased algorithms are some of the untrustful situations that generate power asymmetries between big techs and educational roles. Perhaps the solution can be enforcing legality while maintaining a "human in the loop" to avoid algorithmic decisions and accept/decline algorithmic results.

2. Laws and Geopolitical Issues

The enthusiasm for integrating Big Data processes, data-based decision-making, data processing, and even international transfer between countries has, in some cases, led to problems of misuse, filtering, and improper access [38]. Regarding legal matters, legality, as of today is very far from regulating emerging technologies. We believe that legality is more corrective than preventive. There are still problems concerning trust and loss of control in managing educational data that are not avoided due to legal loopholes. We are aware that there are as many decrees and regulations (from now on, laws) as there are different jurisdictions, some of them more prone to protecting the citizens PICSDM, including educational roles, such as the GDPR in Europe, some less, such as the USA with very low data regulation. To reach a balance between correction and legal prevention and avoid power asymmetries, we believe that it is necessary to have a technological stack that automates every jurisdiction's legal framework, by default and by design. However, an automated ethic is needed, as well as "human-in-the-loop" with some ethical principles. Because legality as a new ethic only helps to privatize institutions and strengthen monopolies, especially when there is no strong regulation and there is strong pressure from lobbies [39].

In terms of geopolitics, we are in a situation of significant change and high sensitivity. The link between Europe and the USA in the search for a regulation of the international transfer of data, the so-called Privacy Shield [40], has been broken after Max Schrems took his original case against Facebook a step further [41] to the point of banning the sharing of data with US entities subject to surveillance laws by their government [42]. The second Schrems II judgment [43] puts the European Commission in focus on adequacy talks with the United States on Standard Contractual Clauses (SCCs). However, the privatization of adequacy assessment is a fact that is manifested in the round tables "Privacy, globalization and international data transfers: towards a new paradigm after "Schrems II" of the International Conference CPDP 2021 [44].

3. Monopolies, Privatization, and Basic Commodities

EdTech services are becoming a staple. Initial phrases such as "memory are not important if the internet is available" or "all knowledge is in your pocket (referring to the mobile phone)" justify the unnecessary work of certain skills. These premises give monopolistic power to the technological ones that have managed to be dependent in an interconnected world. Besides, two factors are accelerating the privatization of educational technology and therefore centralizing educational data on a few actors. On the one hand, the COVID-19 pandemic [45], [46] has accelerated the adoption of educational and business technology in a forced march towards online services and privatization [47]. On the other hand, Williamson [48] points to "learning loss", a reduction in qualified human capital, as the cause of the diversion of capital to EdTech. In Williamson's words, "De-valuations of national economies from 'learning loss' … were mirrored by massive valuations of EdTech, and efforts to capitalize on more of total global education expenditure".

Privatization is compounded by a lack of regulation in the United States that limits the monopolistic power of technology [49]. Educational technology is being capitalized and centralized in a few actors. As found in a longitudinal survey conducted between 2018 and 2019 [13], the educational tools in-app or web platform format used in schools and universities in Spain come mostly from the United States. Specifically, most of the solutions used are companies accelerated at Imagine K12 (IK12) EdTech accelerator [50]. Y Combinator absorbed IK12 during 2016 and so forming its first specialized vertical in education. One company that IK12 worked with was LearnSprout, an online data insights software startup acquired by Apple [51], demonstrating that educational data analytics and insights are of high interest by big technological companies.

IK12 startups use cloud computing or private servers the run their services. Since 2011 more than 10 IK12 startups analyze data collected from students' interactions meaning that data is stored and treated out of the control of academic institutions. Inside LMS in schools and universities students' can be anonymized and respected for their privacy [13]. This is not the case in these EdTech tools when are developed without following the LEDA framework.

4. Mass Surveillance as the New Normal

Most of EdTEch is not open-sourced. This means that is not publicly available to being audited by anyone. Therefore, educational institutions cannot know what it does, how it stores data, how it processes or analyses data, or with whom it shares educational data. Educational actors are subject to blind trust towards these privately funded, and even publicly funded, tools. The LEDA framework considers open-source solutions that protect and secures whilst legally compliant, and entities behind expose an ethical idiosyncrasy. We find entities like Proton [52] and Cryptpad [53] that in some way meet the above four premises and therefore make it possible to create EdTech under these minimum requirements (BLOE):

- Business ethics
- Law compliance
- Open source
- Encryption

Proton and Cryptad are not the only useful tools to save our digital identity and prevent misuse of our data. There are collecting platforms for this type of tool such as ethical.net, operated by the Center for Applied Ethics Ltd. Its slogan "make ethical the new normal" is a definition of principles that puts it in the fight against the surveillance situation. current extreme. Ethical.net [54] is presented as "a collaborative platform for discovering and sharing ethical product alternatives - whether that means purchasing from a social enterprise, thrift shopping, or learning how to fix your old phone instead of buying a new one.". Unfortunately, this description confirms that we are not in an ethical world and that ethics is the alternative when it should be the other way around. Ethical.net is an entity that embraces different ethical issues, however, the technological resources it makes available consider, if not all four BLOE requirements, at least that of business ethics and law compliance. The normality of being monitored, usurped data and metadata, used by third parties as desired and to be used against us requires to be replaced by new normality both ethically and legally. Thus, the 4 BLOE requirements are a start to start changing the scenario, but not enough.

5. Ethics and Data Privacy

Privacy is almost a new term in our society that raises too many questions. It refers to manage sensitive data of students. The introduction of educational data methodologies, such as Learning Analytics [55], [56], raises a non-trustable context in education. Some authors, such as Drachsler & Greller [11], Pardo & Siemens [32], or "Author et al." [19], [57], reflect on ethical issues and how checklists and principles can be useful to diminish data problems.

III. Principles

The seven principles of the LEDA framework [58] have to be seen as a guide for good practices concerning the treatment of educational data in EdTech with a moral basis to resolve concerns, worries, and mistrust in educational data analysis processes raised in the use of loud computing EdTech. Cloud computing is a useful technical architecture in many respects, but that it requires some technical measures and the motivation to redirect current issues related to the privacy and security of collected data.

Without intending to replace the set of interconnectivity and cloud technologies –quite the opposite- the framework of principles is used to generate or use a variant of EdTech that focuses on "local first" instead of "global first", but also considering a mix of both. It is not at all a variant that confronts local installing against the cloud execution, making them exclusive; we do not exclude the use of the cloud. Our alternative promotes in the first place and as a fundamental premise the installation and execution of local applications that can cut off connectivity to the outside for students and give control of the data to teachers, or even give more control to students of their educational interactions. We foster a new perspective in deploying and executing EdTech services.

Considering the above, LEDA principles are stated as:

1. Legality
2. Transparency, information, and expiration
3. Data control
4. Anonymous transactions
5. Responsibility in the code
6. Interoperability
7. Local first

A. Principle 1: Legality

EdTech has to be legally compliant [58], [59]. However, regulations [60] are not automated within EdTech. If regulations are the ultimate solution, we need to update and automate them. Besides, legality is not able to regulate innovations in technology since:

- It is 20 years behind and does not regulate EdTech consistently and correctly [61]
- It may be too restrictive and not allow technology companies to compete with big technologies, or it may be too lax and harm end-users [61]
- It changes considering interests of governmental roles related or not to technological companies [62]

It is complex for legality to contemplate future problems, so legality only embraces known and past problems, under a limited imaginary. Regulating new technologies becomes difficult when you can't imagine or predict what that technology will look like. Besides, strong regulation means that there can be no technological evolution and therefore it will be impossible for alternatives to compete with the ruling monopolies. With weak or non-existent legality, the current privatization of EdTech and other sectors is happening, which also endangers the privacy and confidentiality of data in educational roles. Technology developments must comply with the law and the principles of privacy, confidentiality, and data security must be automated by default and from within the design of the tool [60]. Moreover, automation must be flexible in anticipation of changes in laws and regulations.

Concerning legality and personal data, people share data in really regulated environments but do not fully understand the real consequences:

- Users do not know the laws in depth.
- Users would need an average of 244 hours (30 consecutive working days) [63] to read all privacy clauses and terms of use for each website or application they use daily.

Concerning the above, the concept of privacy paradox stands out, which consists of making decisions via balancing disadvantages and benefits in a situation where people's behaviors do not agree with the principles and values towards private data. However, authors deny the existence of the privacy paradox [64] arguing that people act in the face of the risk presented in each situation and that from these behaviors cannot be inferred the real assessment they have of privacy of your data. People accept legal clauses that otherwise would not be able to operate normally. Therefore, forcing legal protection is beyond the reach of the people themselves, thus reinforcing the argument of the need for automation that makes it possible.

B. Principle 2: Transparency, Information, and Expiration

All educational roles need to be informed of all aspects of data storage and treating. Students should know, among other aspects, what data and metadata will be collected, how long will be stored, who will have access, or which rights will have if technology jurisdiction is different from student jurisdiction. Legal documents are not suitable to understand those such things [64], [65], [66], however, icons can be useful as stated in recital 60 of the GDPR [60].

C. Principle 3: Data Control

Data control refers to:

- Understand how their data are curated, accessed, and shared, under what protocols
- Grant data usage permissions
- Manage data in terms of storage and transfer

Technologists need to understand that to gain trust and be respectful of the digital identity of educational roles, they need to develop solutions that do not require data or only those that are necessary to operate. Therefore, EdTech tools should be considered as mere interfaces with which to interact with local data. These interfaces, as before, might be available from cloud computing in web format or even on student devices. However, this paradigm shift where personal data and metadata reside locally would reduce the excess data transferred and decrease the actors involved.

D. Principle 4: Anonymous Transactions

Data have to be disclosed at will to those who agree with it and only to those who have access to it. Understand and personalize teaching and learning, may need not anonymized data, and this may lay a paradox. However, the act to reveal information to consent people and keep anonymity to others does not suppose a paradox, it supposes privacy by default.

Compliant technology with this LEDA principle uses by default and by design:

- Secure protocols such as SSH, SSL or TLS
- Encryption methods such as Asymmetric Encryption Method or specific encryption algorithms, e.g., AES
- VPN protocols such as OpenVPN

E. Principle 5: Responsibility in the Code

This LEDA principle prioritizes open-source solutions to be evaluated publicly. Responsibility in the code means debating and making questions about ethics and legality in the code [67]:

- What would developers do if asked to write code for an unethical purpose?
- How would developers report unethical code?
- Do developers have an obligation to consider the ethical implications of their code?
- Who is ultimately most responsible for code that accomplishes something unethical?

F. Principle 6: Interoperability

Interoperability protocols, models, techniques, and methodologies already exist. LTI is one of them, and some LMSs, such as Moodle, already carry this built-in technology to interoperate with external tools out of a desire for privacy and data security. However, interoperability alone does not guarantee confidentiality but allows data to be moved freely and at will under established, validated, and functional protocols in trusted environments.

G. Principle 7: Local First

1. Zero Distance

Technology that complies with the "local first" principle is such technology that reduces data transfer distance between devices. It implies that in the development and integration of EdTech, closer technologies are considered first, instead of distant ones such as cloud computing. This makes a shift in considering which technology adopt in academic organizations and also in developing EdTech solutions. "Local first" principle conceives EdTech tools as mere interfaces that process local data. This raises the question; how can the distance concept be used to develop and deploy EdTech tools as mere interfaces?

The "local first" approach does not exclude any other technology, it simply prioritizes how to solve different situations. Cloud computing offers huge opportunities and risks as well. The "local first" approach also offers huge opportunities in a new way of solving issues related to educational data. However, this "local first" new way of doing has to be evaluated. It must be said that an absolute local approach does not solve all problems, even can make solutions complicated and unusable in user experience sense. Therefore, in this "local first" approach, hybrid solutions with cloud computing are not excluded.

2. Protocols and Solutions

The EdTech that uses local storage in users' devices and no data transfer by default is set the maximum value; therefore, this EdTech complies with the most important LEDA principle "local first". Thus, one way to increase confidence in educational settings is to use zero-distance technology. Zero-distance technology is about:

- Communicating, as shown in Fig. 1 and Fig. 2, using near possibilities such as WI-FI, Bluetooth, or NFC. Technologies such as scanning images through mobile cameras would be allowed, e.g., the use of encrypted QR codes to read students' answers to questionnaires.
- Local storage technologies such as Web Storage API, IndexedDB API, or even Cache API for web browsers. These web browser available technologies are specially identified as zero-distance technologies because they enhance privacy and security in the development of EdTech.

As seen in Fig. 1 and Fig. 2, the "local first" concept allows the use of local server applications to keep data safe in rest inside educational organizations. Zero-distance technology enhances data control without allowing third-party actors to catch data generated in learning processes mediated by digital technology solutions. Teachers and students communicate with each other without transfer data to third-party servers. When zero-distance technology is used in EdTech, data is being transferred to company servers. An example of EdTech solutions that partially use "local first" technologies is Plickers [68]. Teachers scan QR codes held by students, but answers and students' data are uploaded to Plickers' data servers, breaking the data control principle.

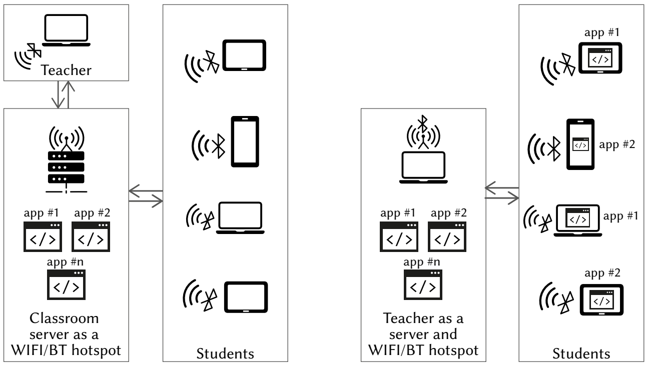


Fig. 1. The use of a WIFI or Bluetooth communication between educational roles where an in-class device (e.g., computer) is used as a server to keep data locally without transfers to third parties.

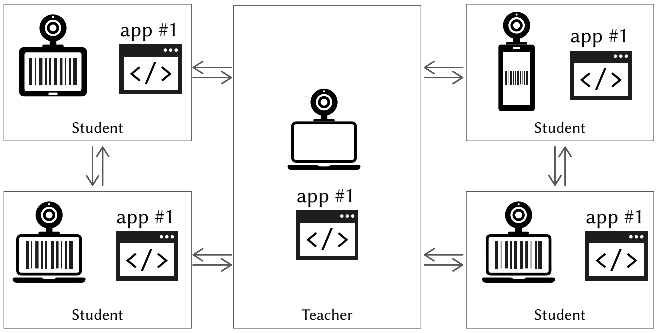


Fig. 2. The use of QR codes or Codebars to share information between educational roles both in-class and out-class to preserve data locally without express transfers to third parties.

3. Synchronization

"Local first" enables synchronization at will. EdTech that uses cloud computing incorporates automatic algorithms that continually communicate devices and servers sending data by default and with no settings to change how, what, and when to synchronize. Data is transferred automatically and constantly to remote company private servers. In the imaginary of LEDA framework, EdTech tools are mere interfaces and data transfer is one-way from student to teacher or student to LMS where students' data can be accessed.

To avoid continuous data streaming to remote servers, synchronization options are enabled but not activated by default. Automation simplifies processes but synchronizes with a local server or personal device when desired allows data management and control. Notwithstanding the evidence that reducing data transfer to cloud computing is the most convenient scenario, not all applications can work completely locally, and some data would be transferred to third-party servers to ensure the proper functioning of the solutions. Regarding "Author et al." that every single data point is fragile, some data points are inevitably shared but not necessarily those that identify persons. Therefore, makes sense to deploy a Personal Data Broker [69] to manage data sharing.

IV. New Looks at "Local First"

Looking into a "local first" approach means reducing two key elements in EdTech:

- Devices, in their broad aspect: servers, routers, hard drives, etc.
- Intermediaries, any actor that can access data from educational roles

Different actors are involved in the use of EdTech, such as classmates, teachers, system administrators of the educational institution, different people from different departments of the educational institution, staff associated with the educational services of third parties, and even third-party staff who have access to data that the educational services share for its proper functioning. Similarly, data navigates complex circuitry through a good set of devices such as proxy servers, switches, or routers where data is likely to be stored.

The EdTech scenario is a complex one and it is not easy to find a unique solution to the problem of fragility in PICSDM. We believe that this scenario can only be resolved through the adoption of standards, good practices, and forced automation under public audits. However, these standards should be based on principles that we believe the LEDA framework is a good starting point for.

The teacher must be able to view the student's data to address, communicate and tutor them, even if this data does not show the student's actual identity [12]. It should also be able Of course the teacher should be able to see academic data as results of questionnaires and other results of assessment instruments. The data that is generated on the student's device should go out. We have proposed possible solutions based on communication technologies within the classroom and others via QR code in case of working online. However, in some environments, there is no presence and QR codes do not solve all the cases. In this sense, given that the student's data will be transferred to another device or storage, principles 3, 4, and 5 take force so that the student has control over data transfers (e.g., acceptance), remain anonymous under the use of interoperable technologies.

A. Personal Data Record Store and Personal Data Brooker

Applying the "local first" principle in its entirety means that the student or any educational role can grant access to the data at will even without the person asking for it despite knowing the real identity. It is not even necessary for a teacher to know the identity of their students, as it will assess evidence of learning and not personalities. It would therefore make sense that in an LMS in absolutely online learning even the personal identifying information of each student is not needed. In any case, some scenarios require data to be transferred out of the student's device for the smooth running of the learning process.

We believe that the use of local storage that stores student interactions in sync with microservices can facilitate the scenario discussed. Fig. 3 shows how this scenario could be and what we are currently working on as a line of research for the LEDA framework:

- Apps as mere interfaces: EdTech tools running locally in the browser or as standalone applications without transferring data to the cloud or transferring anonymized data.
- Personal Data Record Storage (PDRS): local service that collects each interaction with EdTech tools synchronizing with a trusted and encrypted storage. This service installed in the device will have a proxy function between local data and the cloud.
- Personal Data Broker (PDB): some authors of this work worked on this concept [69] as a space managed by the student, or educational role, which connects with third party services such as LMS of the educational institution and share what is needed and at the right time. The stored data would be encrypted and served open or anonymized depending on the occasion. This PDB works as a microservice in which to make requests when needed.
- LMS: Although in Fig. 3 the PDB is connected to the LMS database, in reality, and depending on the situation, the data should not be stored. For instance, the teacher may have access to the student's

email, but the LMS system administrator does not need to have access to it. Other data will be interesting to store to streamline management, such as grades, however, these could be associated with a user identifier, even random, which allows the identity of the student to be anonymized.

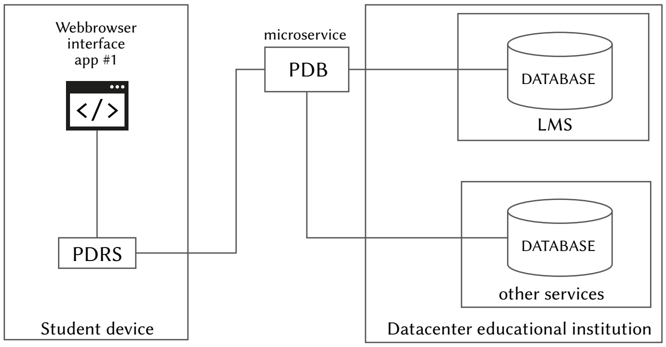


Fig. 3. Use of EdTech tools as interfaces that saves data locally in the Personal Data Record Storage synchronized with the service Personal Data Broker which serves data to educational services if needed.

Microservices such as the PRDS-PDB binomial open up new ways of doing things where data control lies with the student and inquiries can even be anonymous.

V. LEDA Developments

Developing educational tools and solutions following the LEDA framework is not complex, it is a matter of will. However, when designing some tools based on proximity technologies, we realize that usability can be very inconvenient. Cloud Computing makes the difficult easy by automating tedious manual processes. Instead, emulating the same functional behavior in the cloud on the premises becomes something manually if you want to ensure absolute control while using the solution. We present three examples that show how local technologies can add control to both data and processes without the user experience being affected, and how in some cases complexity and inconvenience in using the solutions are added if there are no automatisms. Therefore, the use of local, or zero-distance, technologies require a balance between:

- Complexity
- Automation
- Control

Balancing between one feature or another will depend on the configuration options set by the user. Therefore, the ethics of personalization via configuration options must be defined and implemented, without dark patterns that lead to confusion or wrong decisions.

A. QR Codes

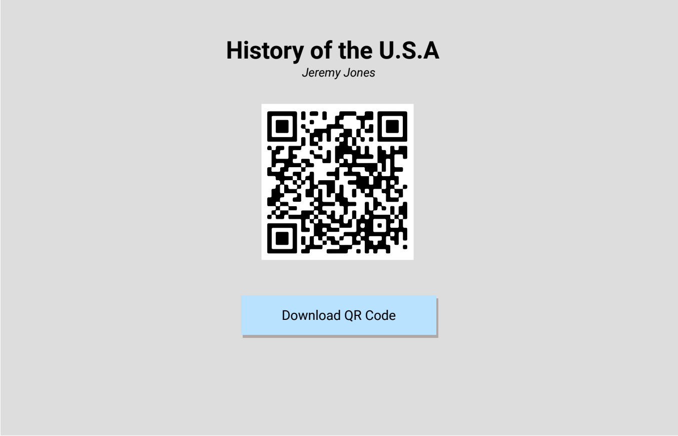
One of the proposals defined in the "local first" principle is implemented, consisting of reducing as many devices as possible using QR codes. In this sense, servers are removed, and an application is designed with the following requirements:

- Teachers should be able to rearrange assessment forms with correct/incorrect questions and answers
- Teachers must be able to import students' answers
- Students must be able to receive the forms
- Students should be able to answer questions and send them to the teacher
- Enable communication between teachers and students via QR codes, either scanned or sent through a conventional communication channel such as email or Bluetooth

This mode of operation allows you to remove proxies and make sharing QR codes immediate by performing a scan. In the worst cases, QR codes can be shared via Bluetooth or other conventional services such as email or flash drive. Sending student responses to the teacher will be done by applying encryption to enable secure communication.

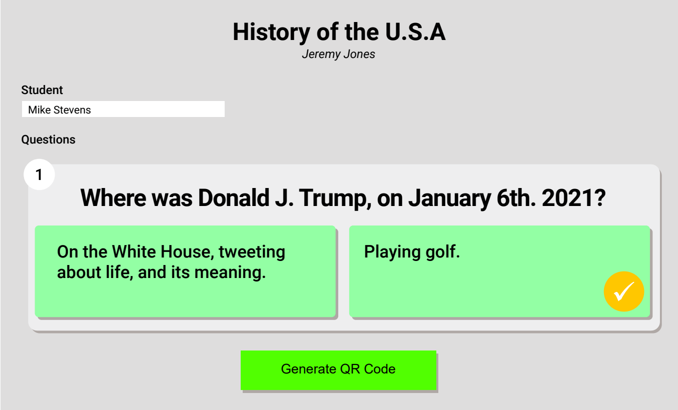
Teachers can enable a QR code repository so that other teachers can scan or import the QR code and dump the content to their devices. Therefore, both student forms and responses will be stored locally on the teacher's device; students will be able to keep their answers on their devices. Fig. 4 shows the teaching version with the on-screen QR code generated from a different question-and-answer form. A button allows you to download the QR and share it if it cannot be scanned. Fig. 5 shows the study version where once the questions have been answered, the QR code can be generated to be scanned by the teacher or sent with the encrypted information.

This solution, however, involves some problems related to the length of the content. Content capacity may require different QR codes to be generated. Therefore, reading by scanning or importing an image can be tedious depending on the number of QRs generated.


History of the U.S.A

Jeremy Jones

Fig. 4. QR code generated from questions and answers introduced by teacher in her device.


History of the U.S.A

Jeremy Jones

Student: Mike Stevens

Questions

1 Where was Donald J. Trump, on January 6th. 2021?

On the White House, tweeting about life, and its meaning.

Playing golf. ✓

Fig. 5. Response of student ready to be qrcoded and scanned or send to teacher.

B. Bluetooth

Other proximity technologies such as Bluetooth connectivity are less cumbersome for users. We developed the Bpoll mobile app. It also consists of creating forms by teachers and responding to students. However, Bluetooth technology makes communication within the classroom easier:

- QR codes are not shared or scanned
- The feeling is like being in a traditional cloud computing application
- Connectivity is local and data is transferred bidirectionally between teacher and student

In this type of technology, it is required to connect the devices expressly. This procedure is partly a control mechanism and partly an additional step that improves the user experience. However, control involves configuration, and configuration leads to a better understanding of technology and awareness of the dangers associated with data. Therefore, the execution process for each role within Bpoll is:

- The teacher creates the forms in the Teacher section
- Teachers warn of the availability of their device via "Advertise device"
- Students are asked to connect to the teacher's device to establish a connection
- Students access the forms and answer the questions presented

Fig. 6 and Fig. 7 show the main screen displayed by the teacher and the connectivity request screen for the student. The process is not far from a cloud computing application except for the pairing step between devices. However, with Bluetooth technology, this process could be automated and pairing reduced to just one and the first evaluation.

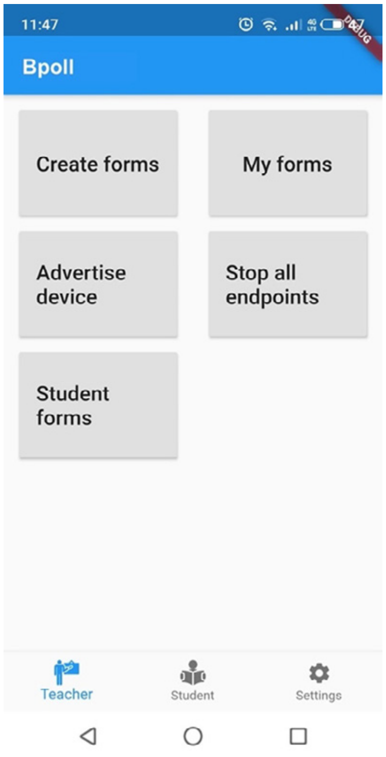
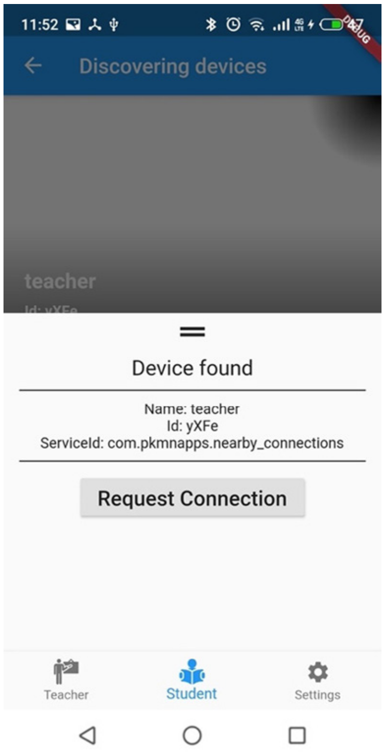


Fig. 6. Bpoll's teacher screen with forms management and connectivity options.

The limitations of using only Bluetooth technology make it usable only in physical spaces due to the limited range. A hybrid technology between Bluetooth and QR codes could bridge these distances given the very limitations mentioned above about QR codes. It is a line of research to resolve these limitations via PRDS and PDB.



Name: teacher
Id: yXFe
ServiceId: com.pkmnapps.nearby_connections

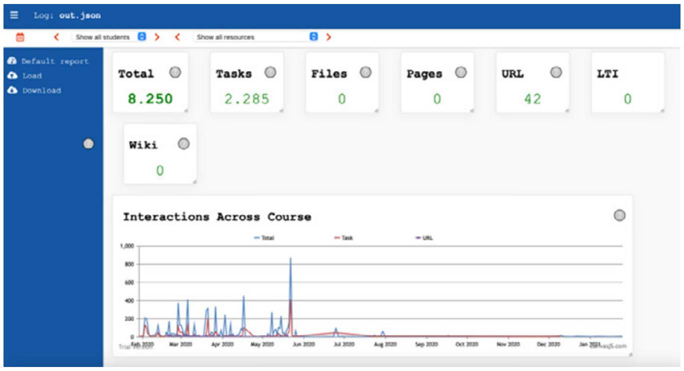
Fig. 7. Students in Bpoll pairing screen with teacher device.

C. Local Analytics

Following the premise of working with local data, we have developed a local execution application, with web technologies available on any personal computer, which allows the analysis of educational data from LMS exports. The development will show that it is possible:

- Distribute the analysis of educational data at the teacher level
- Work on the analysis of local educational data without the need for cloud computing
- In certain environments with a computing power that a browser and a personal computer, including a smartphone, can provide, it is sufficient to perform data analysis.

The tool is called Javascript Moodle Learning Analytics because it is supported by the Javascript Learning Analytics (JSLA) library developed ad hoc and adaptable to any data scheme. Moodle has served as a pilot LMS to use JSLA, build the JSMLA tool (view Fig. 8), and be able to view student interactions from an export of the interaction log.



Log: out.json

Total 8.250
Tasks 2.285
Files 0
Pages 0
URL 42
LTI 0
Wiki 0

Fig. 8. JSMLA dashboard autogenerated by default.

The limitations are given by the computing power of personal devices concerning the amount of data that can be generated in an online class. However, the tool analyzes logs of interactions that are not large in a standard environment. We will soon publish in more detail this open-source web library and all its possibilities.

VI. Research Lines

The LEDA principles open different fields to research on considering local technologies, encryption algorithms, legal considerations, or even ethical approaches. Here we present some research lines as a result of the seven principles:

- Provide an EdTech scoring platform concerning the principles of the framework. Distance between devices and data storage allows a score to be generated and also score EdTech solutions.
- Critically evaluate developments of technological solutions that meet the principles of the framework.
- Fathom the possibilities of the "Personal Data Record Store" and "Personal Data Broker" proposals to encourage the adoption of local technology.
- Detect educational community awareness of the risks related to data and cloud computing, as well as disseminate the seven principles of the LEDA framework as an attenuator and a solver.
- Develop solutions that include by design and by default the directed synchronization, with the possibility of automating the process.

VII. Conclusion

La Salle finds itself in the process of a methodological and technological transformation as well as confronting data issues. La Salle's situation is a very common situation on the educational stage. However, many institutions rely on legality to implement institutional changes that favor pedagogical aspects but also resource reduction in an attempt to balance. Political forces and private capitalism are pushing for the adoption of cloud technologies in an attempt to privatize education. However, when legality is branded as the new ethic, everything is worthwhile, and education is lost. We need technological initiatives and models that eliminate power asymmetries, give control to educational roles and institutions, as well as technologies that automate legality and ethics by omission and from design.

This work sets a framework of data protection and privacy principles to provide trust and confidence to minimize or solve data problems related to EdTech, Data Analytics, and Cloud Computing. This paper sets out the LEDA framework and its seven principles of legality, transparency, data control, anonymous transactions, code responsibility, interoperability, and "local first". This framework has been named LEDA after its seventh principle "local first.". This framework should be considered in the development and adoption of EdTech that collect, store, manage and analyze educational data.

The first six principles set the background to enable the seventh "local first" principle as the most impactful of all in terms of data confidentiality and privacy. The principle "local first" adds new perspectives to adopt ethical EdTech solutions and understand EdTech as interfaces. Data proxies such as personal data record storage enable data control to students allowing access under self-consideration and not by the EdTech tool. The synchronization with personal data brokers inside the LMS facilitates local data management and control to students.

In the framework, distance between devices and EdTech is a key factor to preserve educational data. The more the data is far away, the more perils arise. Zero distance is considered the more adequate scenario. Data proximity is therefore essential to ensure non-leakage, non-misuse, non-prohibited, or inappropriate storage, and non-processing without permission of any data generated in the interaction with the solution.

Considering the limitations of the framework, we envision different problems or educational situations to solve where zero distance makes the solution complex and unusable, considering the use of local-cloud hybrid developments. Local technologies can difficult the user experience and the teaching and learning processes, and even make all educational processes harder to execute. The comfort of cloud computing disappears and can be considered a stopper of local technologies adoption.

The LEDA framework allows institutions to consider issues related to the processing of student data in data-driven decision-making. At the same time, it offers a series of principles of action that minimize or even solve some of the problems present in the adoption of educational technology. These principles offer a change of perception that does not eliminate but ultimately relegates the integration of technologies in Cloud Computing by local technologies.

Acknowledgment

The research that has given rise to these results has been carried out through funds from the Secretariat of Universities and Research of the Department of Business and Knowledge of the Generalitat de Catalunya and Ramon Llull University, with the project 2020-URL-Proj-058.

We acknowledge LEDA solutions development to Àngel Farré for developing Bpoll app, Lluís Masdéu and Guillermo Serraclara for working on the development of QRforms, and Pablo Gómez, Sandra Cea y Nicole Jiménez for UX enhancement of JSMLA.

References

[1] F. J. García-Peñalvo *et al.*, "Mirando hacia el futuro: Ecosistemas tecnológicos de aprendizaje basados en servicios Looking into the future: Learning services-based technological ecosystems," in *La Sociedad del Aprendizaje. Actas del III Congreso Internacional sobre Aprendizaje, Innovación y Competitividad*, Madrid, Spain, 2015, pp. 553–558.

[2] J. O. Islas-Carmona, "El prosumidor. El actor comunicativo de la sociedad de la ubicuidad," *Palabra Clave*, vol. 11, no. 1, pp. 29–39, 2008.

[3] P. Long and G. Siemens. "Penetrating the Fog: Analytics in Learning and Education." EDUCAUSE Review. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education. (accessed Jul. 24, 2018).

[4] D. Amo, M. Alier, M. J. Casan, and M. J. Casañ, "The student's progress snapshot a hybrid text and visual learning analytics dashboard," *International Journal of Engineering Education*, vol. 34, no. 3, pp. 990–1000, 2018.

[5] J. Navarro, A. Zaballos, D. Fonseca, and R. Torres-Kompen, "Master as a Service: A Multidisciplinary Approach to Big Data Teaching," in *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, León, Spain, 2019, pp. 534–538.

[6] C. Bulla, B. Hunshal, and S. Mehta, "Adoption of Cloud Computing in Education System: A Survey," *International Journal of Engineering Science*, vol. 6, no. 6, pp. 6375-6380, 2016.

[7] D. Lupton and B. Williamson, "The datafied child: The dataveillance of children and implications for their rights," *New Media and Society*, vol. 19, no. 5, pp. 780–794, 2017.

[8] R. Mayes, G. Natividad, and J. Spector, "Challenges for Educational Technologists in the 21st Century," *Education Sciences*, vol. 5, no. 3, pp. 221–237, 2015.

[9] D. Amo, M. Alier, F. J. García-Peñalvo, D. Fonseca, and M. J. Casany, "GDPR security and confidentiality compliance in LMS' a problem analysis and engineering solution proposal," in *TEEM'19: Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, León, Spain, 2019, pp. 253–259.

[10] D. Amo, D. Fonseca, M. Alier, F. J. García-Peñalvo, and M. J. Casañ, *Personal data broker instead of blockchain for students' data privacy assurance*, vol. 3. Switzerland: Springer Nature, 2019.

[11] H. Drachsler and W. Greller, “Privacy and analytics: it’s a DELICATE issue a checklist for trusted learning analytics,” in *Proceedings of the sixth international conference on learning analytics & knowledge*, Edinburgh, United Kingdom, 2016, pp. 89–98.

[12] D. Amo, M. Alier, F. J. García-Peñalvo, D. Fonseca, and M. J. Casañ, “Protected users: A moodle plugin to improve confidentiality and privacy support through user aliases,” *Sustainability*, vol. 12, no. 6, p. 2548, 2020.

[13] D. Amo, “Privacidad y gestión de la identidad en procesos de analítica de aprendizaje,” PhD. dissertation, Programa de Doctorado Formación en la Sociedad del Conocimiento, Universidad de Salamanca, Spain, 2020. [Online]. Available: https://repositorio.grial.eu/handle/grial/1951

[14] L. S. D. ARLEP, *NCA, otra manera de hacer escuela*. Madrid, Spain: La Salle ARLEP, 2018.

[15] L. S. D. ARLEP, *NCA, Nuevo Contexto de Aprendizaje*. Madrid, Spain: La Salle ARLEP, 2020.

[16] J. C. Sánchez-Prieto, J. Cruz-Benito, R. Therón, and F. García-Peñalvo, “Assessed by Machines: Development of a TAM-Based Tool to Measure AI-based Assessment Acceptance Among Students,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 4, p. 80, 2020.

[17] D. Amo, M. Alier, F. García-Peñalvo, D. Fonseca, and M. J. Casañ, “Privacidad, seguridad y legalidad en soluciones educativas basadas en Blockchain: Una Revisión Sistemática de la Literatura,” *Revista Iberoamericana de la Educación Digital*, vol. 23, no. 2, pp. 213-236, 2020.

[18] D. Amo, M. Alier, D. Fonseca, F.-J. García-Peñalvo, M. J. Casañ, and J. Navarro, “Evaluation of the importance of ethics, privacy and security in Learning Analytics studies, under the LAK conferences,” in *Actas del V Congreso Internacional Sobre Aprendizaje, Innovación Y Competitividad*, 2019, pp. 343-348.

[19] S. Slade and P. Prinsloo, “Learning Analytics: Ethical Issues and Dilemmas,” *American Behavioral Scientist*, vol. 57, no. 10, pp. 1510–1529, 2013.

[20] B. Williamson, “Decoding ClassDojo: psycho-policy, social-emotional learning and persuasive educational technologies,” *Learning, Media and Technology*, vol. 42, no. 4, pp. 440–453, 2017.

[21] B. Williamson, *Big data in education: The digital future of learning, policy and practice*. London, UK: SAGE Publications Ltd, 2017.

[22] K. Peiró, “ADA en acció: treball de recerca a Catalunya,” in *Intel·ligència artificial. Decisions automatitzades a Catalunya*, Autoritat Catalana de Protecció de Dades and Generalitat de Catalunya, Eds. Barcelona: Autoritat Catalana de Protecció de Dades, 2020, p. 100.

[23] Ø. H. Kaldestad, “New analysis shows how Facebook and Google push users into sharing personal data,” Forbrukerradet, Sentrum, Oslo, 2018. [Online]. Available: https://www.forbrukerradet.no/side/facebook-and-google-manipulate-users-into-sharing-personal-data/

[24] Ø. H. Kaldestad, “New study: Google manipulates users into constant tracking,” Forbrukerradet, Sentrum, Oslo, 2018. [Online]. Available: https://www.forbrukerradet.no/side/google-manipulates-users-into-constant-tracking/

[25] P. Tranberg, “DPA Slams Norwegian Municipalities In Their Use of Google for Education,. DataEthics. https://dataethics.eu/dpa-slams-norwegian-municipalities-in-their-use-of-google-for-education/. (accessed Dec. 20, 2020).

[26] K. Peiró, “És possible acabar amb els biaixos dels algorismes? (1a part).” karmapeiro.com. https://www.karmapeiro.com/2019/06/17/es-possible-acabar-amb-els-biaixos-dels-algoritmes-1a-part/. (accessed Sep. 16, 2020).

[27] F. J. García-Peñalvo, “Learning Analytics as a Breakthrough in Educational Improvement,” in *Radical Solutions and Learning Analytics*, Springer, Singapore, 2020, pp. 1–15.

[28] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, “A reference model for learning analytics,” *International Journal of Technology Enhanced Learning*, vol. 4, no. 5–6, pp. 318–331, 2012.

[29] A. Martínez Monés *et al.*, “Achievements and challenges in learning analytics in Spain: The view of SNOLA,” *Revista Iberoamericana de la Educación Digital*, vol. 23, no. 2, p. 187, 2020.

[30] A. Álvarez-Arana, M. Villamañe-Gironés, and M. Larrañaga-Olagaray, “Mejora de los procesos de evaluación mediante analítica visual del aprendizaje,” *Education in the Knowledge Society*, no. 21, p. 9-13, 2020.

[31] H. Drachsler, “Ethics & Privacy in Learning Analytics - a DELICATE issue,” Learning Analytics Community Exchange. http://www.laceproject.eu/blog/ethics-privacy-in-learning-analytics-a-delicate-issue/. (accessed Sep. 20, 2017).

[32] A. Pardo and G. Siemens, “Ethical and privacy principles for learning analytics,” *British Journal of Educational Technology*, vol. 45, no. 3, pp. 438–450, 2014.

[33] Y.-S. Tsai, P. M. Moreno-Marcos, K. Tammets, K. Kollom, and D. Gašević, “SHEILA policy framework: informing institutional strategies and policy processes of learning analytics,” in *Proceedings of the 8th International Conference on Learning Analytics and Knowledge - LAK ’18*, 2018, pp. 320–329.

[34] L. Amoore, “Why ‘Ditch the algorithm’ is the future of political protest,” The Guardian. https://www.theguardian.com/commentisfree/2020/aug/19/ditch-the-algorithm-generation-students-a-levels-politics. (accessed Sep. 16, 2020).

[35] J. Satisky, “A Duke study recorded thousands of students’ faces. Now they’re being used all over the world,” The Chronicle. https://www.dukechronicle.com/article/2019/06/duke-university-facial-recognition-data-set-study-surveillance-video-students-china-uyghur. (accessed Sep. 16, 2020).

[36] D. MacMillan and N. Anderson, “College admissions officers rank prospective students based on web browsing, family finances and other data,” The Washington Post. https://www.washingtonpost.com/business/2019/10/14/colleges-quietly-rank-prospective-students-based-their-personal-data/. (accessed Sep. 16, 2020).

[37] L. Amoore and R. Raley, “Securing with algorithms: Knowledge, decision, sovereignty,” *Security Dialogue*, vol. 48, no. 1, pp. 3–10, 2017.

[38] M. Grothaus, “Pearson data breach: details of hundreds of thousands of U.S. students hacked,” Fast Company. https://www.fastcompany.com/90384759/pearson-data-breach-details-of-hundreds-of-thousands-of-u-s-students-hacked. (accessed Jan. 8, 2019).

[39] L. Kayali and V. Manancourt, “How Europe’s new privacy rules survived years of negotiations, lobbying and drama,” POLITICO. https://www.politico.eu/article/europe-privacy-rules-survived-years-of-negotiations-lobbying/. (available Feb. 10, 2021).

[40] M. A. Weiss and K. Archick, “U.S.-EU data privacy: From safe harbor to privacy shield,” *The European Union: Challenges and Prospects*. Congressional Research Service, pp. 113–135, 2016.

[41] NYOB, “My Privacy is None of Your Business,” None Of Your Business. https://noyb.eu/en. (accessed Sep. 16, 2020).

[42] P. Petit, “‘Everywhere Surveillance’: Global Surveillance Regimes as Techno-Securitization,” *Science as Culture*, vol. 29, no. 1, pp. 30–56, 2020.

[43] R. Á. Costello, “Schrems II: Everything is Illuminated?,” *European Papers-A Journal on Law and Integration*, vol. 5, no. 2, pp. 1045–1059, 2020.

[44] ElDerecho.com, “CPDP 2021, los retos de las transferencias internacionales de datos tras Schrems II,” ElDerecho.com. https://elderecho.com/cpdp-2021-los-retos-de-las-transferencias-internacionales-de-datos-tras-schrems-ii. (accessed Feb. 9, 2021).

[45] F. J. García-Peñalvo, A. Corell, V. Abella-García, and M. Grande, “Online assessment in higher education in the time of COVID-19,” *Education in the Knowledge Society*, vol. 21, pp. 12-26, 2020.

[46] F. J. García-Peñalvo and A. Corell, “La COVID-19: ¿enzima de transformación digital de la docencia o reflejo de una crisis metodológica y competencial en la educación superior?,” *Campus Virtuales*, vol. 9, no. 2, pp. 83–98, 2020.

[47] B. Williamson and A. Hogan, “Pandemic privatization and digitalization in higher education,” code acts in education. https://codeactsineducation.wordpress.com/2021/02/10/pandemic-privatization-digitalization-higher-education/. (accessed Feb. 10, 2021).

[48] B. Williamson, “De-valuations of national economies,” Message on Twitter. https://twitter.com/BenPatrickWill/status/1360322220132884481?s=20. (accessed Feb. 12, 2021).

[49] C. Doctorow, “How to destroy surveillance capitalism,” OneZero. https://onezero.medium.com/how-to-destroy-surveillance-capitalism-8135e6744d59. (accessed Aug. 26, 2020).

[50] Y Combinator, “Imagine K12,” Imagine K12.http://www.imaginek12.com/. (accessed Sep. 16, 2020).

[51] C. Loizos, “Y Combinator Absorbs Edtech Accelerator Imagine K12, Creating Specialized Vertical,” TechCrunch. https://techcrunch.com/2016/02/10/y-combinator-absorbs-edtech-accelerator-imagine-k12-creating-specialized-vertical/. (accessed Feb. 10, 2016).

[52] ProtonMail, "We have released an open source OpenPGP library - ProtonMail Blog." OpenPGP library. https://protonmail.com/blog/openpgpjs-3-release/. (accessed Sep. 16, 2020).

[53] CryptPad, "CryptPad Analytics & Privacy - What we can't know, what we must know, what we want to know." CryptPad Blog. https://blog.cryptpad.fr/2017/07/07/cryptpad-analytics-what-we-cant-know-what-we-must-know-what-we-want-to-know/. (accessed Sep. 16 2020).

[54] Ethical.net, "Ethical.net." Make ethical the new normal. https://ethical.net/. (accessed Jul. 11, 2019).

[55] C. Lang, G. Siemens, A. Wise, and D. Gasevic, *Handbook of Learning Analytics*. New York, USA: SOLAR, Society for Learning Analytics and Research, 2017.

[56] G. Siemens and R. S. J. d Baker, "Learning analytics and educational data mining: towards communication and collaboration," in *Proceedings of the 2nd international conference on learning analytics and knowledge*, 2012, pp. 252–254.

[57] C. Lang, L. P. Macfadyen, S. Slade, P. Prinsloo, and N. Sclater, "The complexities of developing a personal code of ethics for learning analytics practitioners implications for institutions and the field," in *LAK '18: Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 2018, pp. 436–440.

[58] D. Amo, R. Torres, X. Canaleta, J. Herrero-Martín, C. Rodríguez-Merino, and D. Fonseca, "Seven principles to foster privacy and security in educational tools: Local Educational Data Analytics," in *TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2020, pp. 730-737.

[59] Boletín Oficial del Estado, "LOPDGDD BOE-A-2018-16673," *Boletín Oficial del Estado, núm. 294*, 2018. [Online]. Available: https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673. (accessed Sep. 6, 2019).

[60] EP and the CEU, "Regulation (EU) 2016/679 GDPR." Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679. (accessed Jun. 27, 2019).

[61] A. Pedreño and L. Moreno, *Europa frente a EE.UU. y China. Prevenir el declive en la era de la inteligencia artificial.* Alicante, Spain: Amazon, 2020.

[62] J. P. Carlin and G. M. Graff, "Dawn of the code war: America's battle against Russia, China, and the rising global cyber threat." PublicAffairs. https://www.publicaffairsbooks.com/titles/john-p-carlin/dawn-of-the-code-war/9781541773813/. (accessed Sep. 16, 2020).

[63] A. M. Mcdonald and L. F. Cranor, "The Cost of Reading Privacy Policies," *I/S: A Journal of Law and Policy for the Information Society*, vol. 4, pp. 1–22, 2008.

[64] D. J. Solove, "The Myth of the Privacy Paradox," *SSRN Electronic Journal*, vol. 89, no. 1, pp. 1-51, 2020.

[65] P. Graßl, H. Schraffenberger, F. Z. Borgesius, and M. Buijzen, "Dark and bright patterns in cookie consent requests," *Journal of Digital Social Research*, vol. 3, no. 1, pp. 1-38, 2020.

[66] D. Machuletz and R. Böhme, "Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR," *Proceedings on Privacy Enhancing Technologies*, no. 2, pp. 481–498, 2020.

[67] Stack Overflow, "Stack Overflow Developer Survey Results 2018." Developer Survey Results, 2018. https://insights.stackoverflow.com/survey/2018/. (accessed Jan. 1, 2019).

[68] Plicekers, "Plickers." Website of Plickers educational app. https://get.plickers.com/. (accessed Sep. 16, 2020).

[69] D. Amo, D. Fonseca, M. Alier, F. J. García-Peñalvo, M. J. Casañ, and M. Alsina, "Personal Data Broker: A Solution to Assure Data Privacy in EdTech," in *International Conference on Human-Computer Interaction*, Orlando, USA, 2019, pp. 3–14.

Daniel Amo

PhD in Education Sciences from the University of Salamanca (2020), with two master's degrees in education and educational technology, University Master's Degree in Teacher Training for Compulsory Secondary Education and Baccalaureate, Professional Training and Language Teaching (UNIR 2016) and University Master's Degree in Education and ICT, specialization in Research (UOC 2014). He currently focuses his professional career on University teaching in the Department of Computer Engineering at La Salle, Ramon Lull University, and Research in the established research group GRETEL (Group of REsearch in Technology Enhanced Learning) recognized by the Generalitat de Catalunya within the call 2017 SGR 934. Within GRETEL he coordinates the Educational Data Analytics research line with a special focus on the field of Learning, Feedback and Ethical Analytics. With his thesis dissertation "Privacy and identity management in learning analytics processes" he adds aspects related to Privacy, Identity, Confidentiality and Security of Personal Data, Data and Metadata in the educational context in his research career. Since 2014, it has published more than 30 scientific articles related to Education in areas such as Educational Analytics, Educational Technology, MOOCs, Educational Networks or Privacy and Ethics in Education. He actively participates in scientific congress committees (CISTI'21) and conferences to disseminate to society with knowledge resulting from his professional and personal research. He is the author of the books "Learning Analytics: the narrative of learning through data" (UOC OuterEdu), "Learning analytics: 30 experiences in the classroom with data", and of the Learning Analytics' divulgation blog edulíticas.com.

Paul Prinsloo

Paul Prinsloo is a Research Professor in Open and Distance Learning (ODL) in the College of Economic and Management Sciences, University of South Africa (Unisa). His academic background includes fields as diverse as theology, art history, business management, online learning, and religious studies. Paul is an established researcher and has published numerous articles in the fields of teaching and learning, student success in distance education contexts, learning analytics, and curriculum development. His current research focuses on the collection, analysis and use of student data in learning analytics, graduate supervision and digital identity. Paul was born curious and in trouble. Nothing has changed since then.

Marc Alier

Marc Alier (1971) received an engineering degree in computer science (1996) and a PhD in Sustainability (2009) in the Polytechnical University of Catalonia (UPC). He is an associate professor at UPC and deputy director at ICE http://www.ice.upc.edu. The last 25 years have worked in research and development related to the e-learning industry. Has participated in the development of several LMS, content authoring tools and interoperability standards. Since 2001, has taught software engineering, project management, information systems, and computing ethics at UPC's School of Informatics. Has been director of several master's programs. Has authored more than 120 papers in journals and conference proceedings. Since 2007 produces several podcasts about technology, science and its impact on society as a means of dissemination of his professional and personal research.

David Fonseca

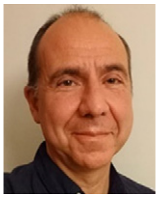
Full Professor (2017) by La Salle Ramon Llull University, currently he is the coordinator of the Group of Research on Technology Enhanced Learning (GRETEL), a recognized research group of Generalitat de Catalunya (from 2014), and coordinator of the Graphic Representation Area in the Architecture Department of La Salle (where he is a teacher and academic tutor). Technical Engineer in Telecommunications (URL – 1998), Master in GIS (Universitat de Girona, 2003), Audiovisual Communication Degree (UOC, 2006), Master in Advanced Studies (URL-2007), Official Master in Information and Knowledge Society (UOC, 2008), PhD in Multimedia by URL (2011), also, he is Autodesk Approved and Certified Instructor from 1998. With extensive experience in project manager (from 2000 to act, he has coordinated more than 50 local, national, and international projects, both technological transfer and research funded projects), he has directed 7 PhD thesis and more than 10 other final degree and master projects. Currently he is serving as program or scientific committee in more than 15 indexed journals and conferences, as well as organizing workshops, special issues and invited sessions in different scientific forums.

Ricardo Torres Kompen

Lecturer, Coordinator of the Degree in International Computer Engineering. Researcher in the area of learning technologies and university professor since 1996. He is a chemical engineer (1991), a Master in Chemical Engineering (2000) and a PhD in Multimedia Engineering from the Polytechnic University of Catalonia–BarcelonaTech (2016). His research focuses on the personalization of learning through the use of multimedia and technology.

Xavier Canaleta

Dr. Xavi Canaleta has a degree in Computer Science from the Facultat d´Informàtica de Barcelona (Universitat Politècnica de Catalunya) and PhD from the Ramon Llull University. He has been developing academic functions in La Salle - Universitat Ramon Llull since 2001. He was Vicedean for Educational Innovation, Coordinator of the Degree in Computer Engineering, and the Master´s Degree in Teacher Training (Technology specialty). He is also professor of Operating Systems (Degree) and of Teaching Innovation and Educational Research (Master). In his academic trajectory he has been professor of different subjects of Engineering degrees (Advanced Operating Systems, Programming, and Introduction to Computers) and of masters (Technology in the Social Context and Practices in Educational Centers). Between 2003 and 2005 he was Coordinator of Computer Services at the Ramon Llull University. As far as research is concerned, he belongs to GRETEL (Group of Research in Technology Enhanced Learning) and has participated in various European projects and has written articles in national and international journals and conferences. His origins have a 14-year experience with teacher and coordinator in Secondary and High School.



Javier Herrero-Martín

PhD in Psychology. Vice Dean of Undergraduate Studies in Early Childhood Education and Primary Education at the La Salle Higher Center for University Studies (Autonomous University of Madrid). Professor at La Salle Campus Madrid, Faculty of Education and Education. Expert in cognitive and language psychology (Complutense University of Madrid). Director of the INAEX-La Salle research group. Member of the District Educational Innovation Team NCA-ARLEP (La Salle, Spain and Portugal).

unir
LA UNIVERSIDAD
EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net