Two-Stage Human Activity Recognition Using 2D-ConvNet

Kamal Kant Verma[1]*, Brij Mohan Singh[2], H. L. Mandoria[3], Prachi Chauhan[4]

[1] Research Scholar, Uttarakhand Technical University, Dehradun (India)
[2] Department of Computer Science and Engineering, College of Engineering Roorkee, Roorkee (India)
[3] Department of Information Technology, G.B Pant University of Agriculture and Technology, Pantnagar (India)
[4] Research Scholar, Department of Information Technology, G. B Pant University of Agriculture and Technology, Pantnagar (India)

Received 25 October 2019 | Accepted 11 March 2020 | Published 24 April 2020

unir
LA UNIVERSIDAD EN INTERNET

Abstract

There is huge requirement of continuous intelligent monitoring system for human activity recognition in various domains like public places, automated teller machines or healthcare sector. Increasing demand of automatic recognition of human activity in these sectors and need to reduce the cost involved in manual surveillance have motivated the research community towards deep learning techniques so that a smart monitoring system for recognition of human activities can be designed and developed. Because of low cost, high resolution and ease of availability of surveillance cameras, the authors developed a new two-stage intelligent framework for detection and recognition of human activity types inside the premises. This paper, introduces a novel framework to recognize single-limb and multi-limb human activities using a Convolution Neural Network. In the first phase single-limb and multi-limb activities are separated. Next, these separated single and multi-limb activities have been recognized using sequence-classification. For training and validation of our framework we have used the UTKinect-Action Dataset having 199 actions sequences performed by 10 users. We have achieved an overall accuracy of 97.88% in real-time recognition of the activity sequences.

Keywords

Activities Recognition, Random Forest, 2D Convolution Neural Network, Intelligent Monitoring System.

DOI: 10.9781/ijimai.2020.04.002

I. Introduction

RECOGNITION of human activities have been a hot research field in computer vision for more than two decades and researchers are still working in this domain due to unavailability of perfect human activity recognition system. Still images give less knowledge for action recognition as compared to the videos. Videos give temporal information as an additional ingredient, which is an important indicator for action recognition. A large number of different activities may be correctly identified depending on the motion component found in videos.

Action recognition is an active ingredient of many applications such as automatic video surveillance [2]-[6], object detection and tracking [7], video retrieval [8], etc. Other applications are strongly connected to the activities and actions recognitions, like human motion analysis [9]-[15], analysis of dynamic scene activities [16], classification of human actions [17], or understanding human behavior [18]. Human activity recognition comprises various steps, which define the features that represent low level activities. The activities of interest and their details may vary depending on the applications. For example, from the last few years Automatic Teller Machine (ATM) has become one of the prime facilities for cash dispense, cash withdrawal, balance enquires, etc. For this reason ATM has become an unsafe site and if the security issues of ATM are concern then, it requires an intelligent video surveillance system that not only captures the scene information at the time of abnormality, but also recognizes single and multi-limb human abnormal activities so that the intelligent system could warn to the security-in-charge in real time and the corrective action could be taken at the time when the abnormal activity happens either by single or multi human limb.

A basic model of human activity recognition in video frame sequences consists of mainly two levels. In the first level, handcrafted features have been extracted from raw input data and in the second level a classifier model is built depending on these features. Here some of the most frequently used feature detectors for human activity recognition have been discussed, which include Histogram of optical flow (HOF), Spatial-Temporal Interest Points (STIP), Histogram of Oriented Gradients (HOG), and dense trajectories [43] etc. However, the extraction of these features is a really difficult and time consuming process as well as it is challenging to know which kind of feature is relevant to the problem because feature selection varies from problem to problem in real time. Therefore a deep learning based model has been proposed and discussed in below section to attend the demand for handcrafted features and reduce the complexity of this process.

At a recent time, deep learning has arisen as a group of deep architecture based learning models that render high-level abstraction of data. A deep learning model is a systematic presentation of multiple

* Corresponding author.
E-mail address: kkv.verma@gmail.com

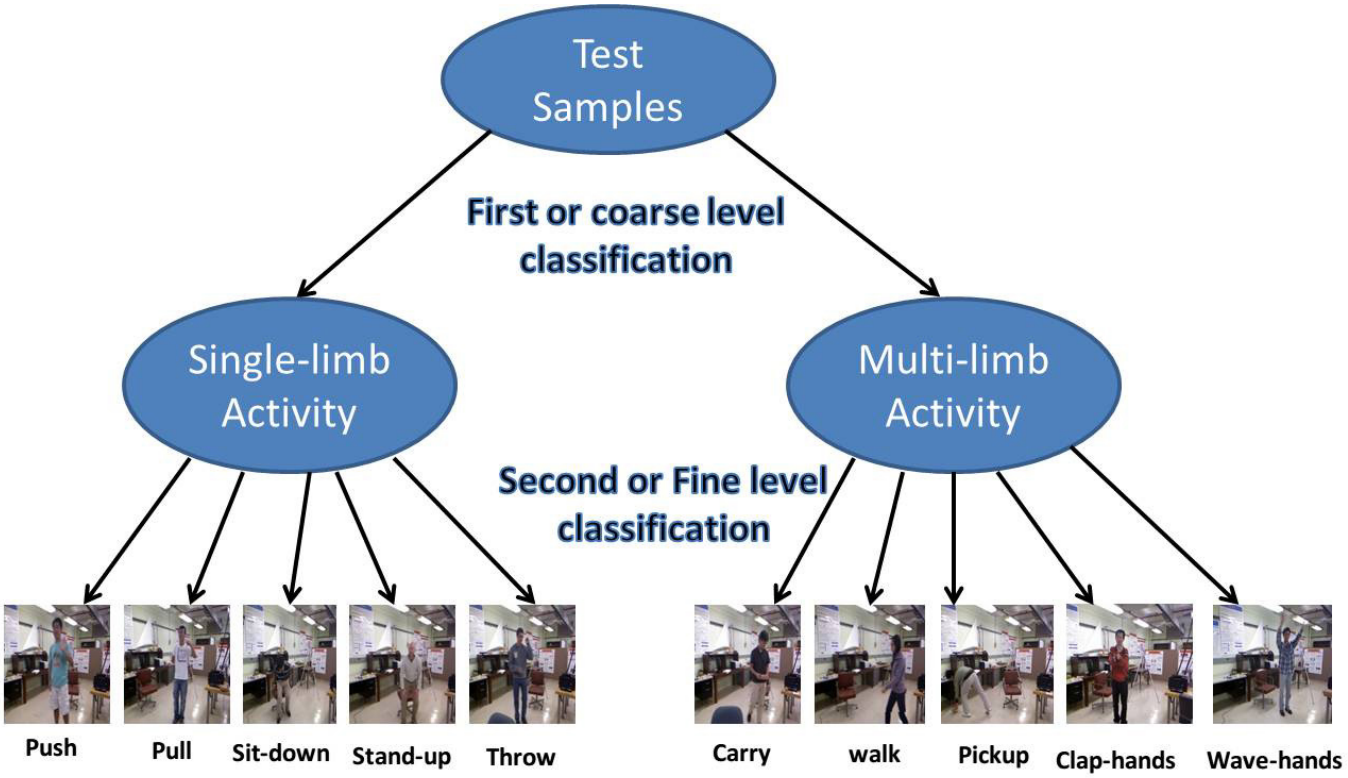


Fig. 1. Two-Stage Classification of human single-limb and multi-limb activities.

layers that are organized for automatically learning of features. Moreover, every layer in the deep architecture model receives output from the previous layer and implements nonlinear transformation such that the input data are transformed into systematic order of low level features to more advanced level features. The most common types of deep learning models are convolutional neural networks, recurrent neural networks, auto encoders, deep belief networks etc. For the labeled input data the deep learning model is trained based on supervised learning and, in case of unlabeled data, the deep architecture model is trained via unsupervised learning. Due to its outstanding performance in various areas like bio signal recognition, gesture recognition, computer vision, bio-informatics, etc. it could be fully deployed in human activity recognition.

II. Related Work

Action recognition in video frames, images sequences or in still images have become a hot research area over the past several years. Since it is not possible to discuss complete literature of action recognition, hence major focus has been given on action recognition in (a) RGB video frames (b) depth frames using deep neural network. Due to the increasing demand of computer vision, latest research work has shifted towards applying convolution neural networks (CNNs) for activity recognition because it is able to learn spatio-temporal information [20], [21], [22] from the videos. Li et al. in [23] used CNN for recognition of human activity captured using a smart phones data set. However activity recognition is not the only field where convolution neural network has achieved wonderful results but it outperforms in various areas such as human facial recognition [24], image recognition [25]-[27] and human pose estimation [28].

The first work on MSR3DAction dataset was given by L. Xia et al. in [29], a histogram of 3D joints (HOJ3D) has been computed and redirected by linear discriminant analysis and divided in k different visual words and temporal evaluations of these visual words are formed using HMM. L. Xia et al. validated their approach on MSR UT kinect-Action 3D joints dataset and achieved 90.92 % accuracy. The recently proposed work in [30] used CNNs, Long Short-Term Memory (LSTM) units and a temporal-wise attention model for action recognition. To learn visual features using CNNs [25] have given the benefit over hand-crafted features for recognition in still images [31]-[33] moreover it overcomes the limitation of the manual feature extraction process. Modified CNNs for recognition of activities in video frames was suggested in various approaches [20], [22], [26], [34]-[39]. A couple of methods used single video frames with spatial features [20], [26], [34]. Multi-channel inputs to 2-dimensional convolution neural networks have also been used in [20], [26], [37], [39]. In [26] author divided the temporal and spatial video parts by using the optical flow method of RGB frames. Each of the separated parts was placed into different deep convolution neural network for learning spatial and temporal features of appearance and movement of object in a frame. Moreover, activity recognition can be performed not only by finding spatial temporal information but a state-of-the-art video representation in [40]-[42] also uses dense point trajectories. The first work was given in [43], which uses dense points of each frame and follows these points based on displacement information after finding a dense optical flow field. The method proposed in [43] was validated on four bench mark datasets such as KTH, YouTube, Hollywood2, and UCF-sports. The better implementation of the approach based on trajectory was given in the Motion based Histogram [44] which was calculated on vertical and horizontal parts of optical flow [50].

However the recent CNN approaches use 2D-convolution architecture of the video which allows learning the shift-invariant representations of the image scene. Meanwhile 2D-convolution is unable to incorporate the depth volume of video which vary with time and is also an important ingredient for activity recognition from the beginning to the end of the activity. 3D-Convolution addresses this issue and incorporates spatio-temporal information of videos and provides a real extension of 2D convolution. A different way to deal with spatio-temporal features is

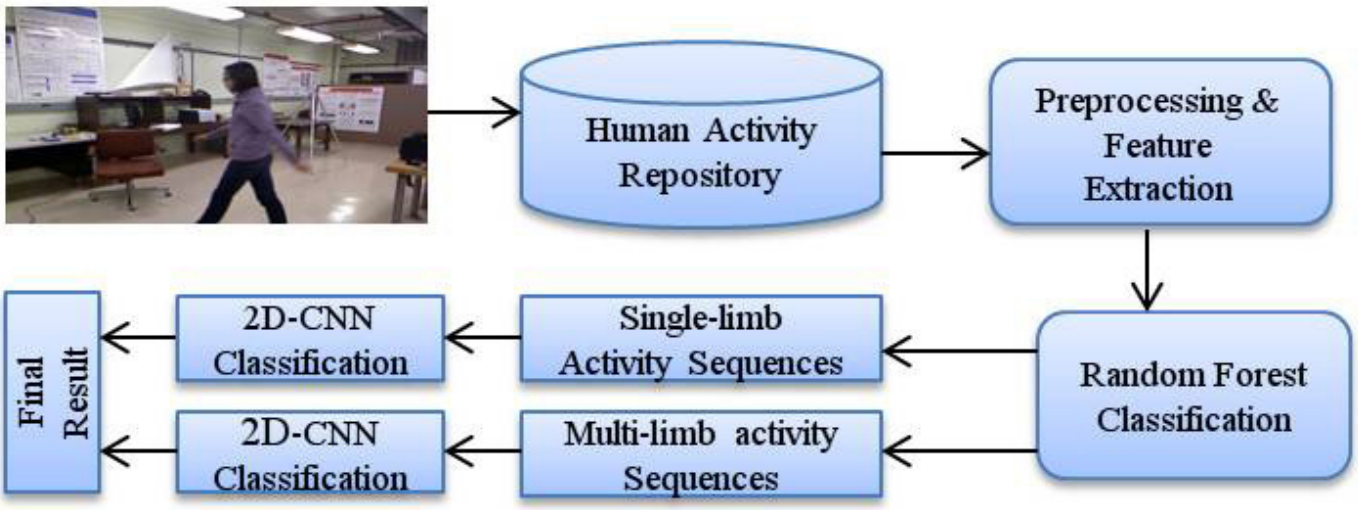


Fig. 2. Flow diagram of the proposed framework.

suggested in [45], where the author factorized original 3D convolution into 2D spatial in the lower layer followed by one-dimensional temporal convolution in the upper layer. The proposed framework was tested on two bench marked datasets (UCF-101 and HMDB-51) and outperformed over existing CNN based methods.

Liu et al. [46] proposed an approach for human activity recognition using coupled hidden conditional random fields (CRF) by combining RGB and depth data together rather than CRF because it had the limitation that it cannot capture the intermediate hidden state using variables. Liu et al. implemented their method on three datasets named DHA, UT kinect-Action 3D and TJU, which produced 95.9%, 92% and 92.5 % accuracies respectively. Zhao et al.[47] proposed a multi-model architecture using 3D Spatio-temporal CNN and SVM using raw depth sequences, depth motion map and human 3D skeleton body joints for recognition of human actions, the proposed approach was validated on UTKinect-Action 3D dataset and MSR-Action 3D dataset and gave 94.15% and 97.29% accuracies respectively. Arrate et al. in [48] suggested 3D geometry of different human parts by taking translation and rotation in the 3D space and tested the approach on UTKinect-Action 3D dataset, achieving 97.02% of accuracy. Siirtola et al. in [51] discussed the personal human activity recognition model using incremental learning and smartphone sensors. Authors in [51] have also discussed that how human activity recognition system has changed since 2012, and how this human activity recognition method can be used in healthcare application. Jalal et al. in [52] suggested a novel framework for human behavior modeling using 3D human body posture captured from Kinect sensor based on clue parameters. In [52] authors extracted human silhouettes from noisy data and tracked body joints values by considering spatio-temporal, motion information and frame differentiation, then angular direction, invariant feature and spatio-temporal velocity features have been extracted in order to find the clue parameters, at last clue parameters are mapped into code-words and recognize the human behavior using advanced Hidden Markov Model (HMM), the proposed approach is tested on three benchmark depth datasets: IM-DailyDepthActivity, SRDailyActivity3D and MSRAction3D, and got 68.4%, 91.2% and 92.9% accuracies respectively.

Simonyan et al. [26] used two convolution neural networks for two streams, one for spatial features and another for temporal features and validated the proposed system on different bench-mark datasets, UCF-101 and HMDB-51. The limitation of this method [26] is that they applied a complete dataset for spatial stream and temporal stream, both for 101 classes' data in UCF-101 dataset and 51 class's data for HMDB-51 and achieved 88% and 59.4 % accuracies, respectively. There may be a scope of research to improve the results in terms of recognition accuracies by applying our two stage methodologies. Similarly we can apply our proposed approach to any other challenging large scale action recognition dataset.

III. Motivation

In our proposed work, we have used a novel multi-stage classification framework based on the color information retrieved from a RGB camera in an intelligent video surveillance system. In our work we have used two-stage classification. The advantage of two-stage classification is that we can handle a large complex problem in a better way in real time since it is difficult to train the network when the number of classes are high.

Therefore in this approach we have divided the classes into two categories at subsequent level. In two-stage classification the first stage is also known as coarse level classification and the second stage is called fine level classification. The two stage classification is given in Fig. 1, as Wang et al. [30] used CNN-LSTM based attention model for human action recognition on UCF-101 dataset and achieved 84.10% accuracy. Similarly Karpathy et al. [20] used CNN on sports dataset for sport action recognition and got 63.3% accuracy. Both, Wang et al [30] and Karpathy et al. [20] used complete dataset only at once and performed classification, which was the limitation of the work given in [20] and [30]. If our proposed two-stage approach had been used on the datasets mentioned in [30] and [20], we could get better results.

However, our approach divides the number of classes into multi-levels and reduces the loss first at initial level, and then at subsequent levels, overcoming the limitation of work given in [20] and [30] in which whole dataset was taken at once, hence by splitting the dataset into a multi-level scenario we applied the proposed technique and achieved good results in terms of accuracy compared to the state of art literature.

Our proposed method may be applied to large scale human action datasets such as UCF-101 [22], [34], [36]-[37], HMDB-51[26], [38], [39], [41]-[42] which divides the whole dataset into a subsequent level of hierarchy in downward direction, which leads to the reduction in losses at each level. Our proposed method is applied and validated on UTKinect-Action 3D dataset, giving promising result and advances as compared to the previous literature.

In this work, a 2D-Convolution Neural Network has been used due to its excellent performance on object detection, activity classification and recognition. It has the power of automatically learning the features from the input video data. The work recently published in [19] used a stack of 3D-CNN on spatio- temporal activation re-projection (STAR-Net) using RGB information. Most of the previous approaches have utilized hand crafted features, which is a really time consuming process. Therefore this work proposed a two-stage-DNN based model to recognize the human activities. In the first stage we divided all activity types into two categories based on human limbs which are human single limb and human multi limb activities. In the second stage two 2D-Convolutions Neural Network have been used for activity recognition.

From the experimental point of view, UTKinect-Action3D Dataset has been used which contains the following ten human activity classes types: wave hands, pull, walk, sit down, push, stand up, throw, pick up, carry, and clap hands. Based on our proposed approach we have divided sit down, standup, pull, push and throw into human single limb activities category, and walk, pickup, carry, wave hands, clap hands, into human multi-limb activities category. Table I shows some of the state-of-art techniques, datasets used, accuracies, number of classes in datasets used alongside with their applications.

In summary, the main work of this paper is given as:

- This paper proposes a two-stage activity recognition framework for RGB information. In the first stage human activities are distinguished based on human single and multi-limb categories.
- In the second stage two Deep-CNN models are used to recognize the separated single and multi-limb human activities.

TABLE I. Some Of Art-Of-The-State Approaches With Their Datasets And Accuracy

Name	Approach	Datasets	Accuracy	Classes	Applications
Zhang. et al.[9]	Affine Transformation, SVM	-	-	-	Pedestrian Detection
Wang et al.[30]	CNN, LSTM and Attention Model	UCF-11,UCF-Sports and UCF-101	98.45%,91.9% and 84.10%	11,10,101	Video Action Recognition
Karpathy et al. [20]	CNN	Sports-1M	63.3%	487	Recognition of Sports actions
Taylor et al.[21]	Convolution Gated RBM	KTH action dataset, Hollywood2 dataset	90% and 47.4%	6,12	Human action recognition
Tran et al.[22]	3DCNN+SVM	UCF101	52.8%	101	Action Recognition
Xue et al.[23]	CNN	HARUSP dataset	96.1%	6	Sensor based activity
Schroff. et al.[24]	CNN+L2 Normalization +Triplet Loss	YouTube Face DB Dataset	99.63%	-	Face Recognition
Simonyan et al.[26]	Two Stream ConvNet	UCF-101,HMDB-51	88% and 59.4%	101,51	Action Recognition
Xia. et al.[29]	HOJ3D+Lin. Discriminant +HMM	MSR3DAction Dataset	90.92%	10	Human Action Recognition
Girshick. et al.[32]	Region-CNN	ILSVRC2013 detection dataset	47.9%	200	Object detection and region segmentation
Taigman. et al. [33]	Deep CNN	(LFW) and You-tube datasets	97.35% and 91.4%	-	Face Recognition
Donahue. et al. [34]	CNN+LSTM	Face(YTF) UCF-101	68.2%	101	Action Recognition
Xu. et al. [35]	3DCNN	TRECVID 2008 And KTH Dataset	72.9% and 67.52%	3,6	Action Recognition
Wang. et al. [36]	trajectory-pooled deep-convolutional descriptor (TDD)	UCF-101	84.7%	101	Action Recognition
Wang. et al. [37]	GoogleNet+VGG16 based Two Stream ConvNet	UCF-101	91.4%	101	Action Recognition
Bilen. et al. [38]	Dynamic Image + CNN	UCF-101 and HMDB-51	89.1% and 65.2%	101,51	Action Recognition
Feichtenhofer. et al. [39]	Two Stream Network Fusion CNN	UCF-101 and HMDB-51	93.5%and 69.2%	101,51	Action Recognition
Peng. et al. [41]	Stacked Fisher Vector(SFV)	YouTube, J-HMDB and HMDB51dataset datasets	93.77%,69.03% and 66.79%	11,51,21	Action Recognition
Wang. et al. [42]	Dense trajectories using SURF and optical flow Features + RANSAC	Hollywood2, HMDB51,Olympic Sports and UCF051 Dataset	64.3%,57.2%, 91.1% and 91.2%	12,51,16 and 50	Action Recognition
Sun et al.[45]	Factorized Spatio-temporal CNN	UCF-101 and HMDB-51	88.1% and 59.1%	101 and 51	Activity Recognition

The paper is arranged as follows: section II contains the related work on activity recognition in RGB videos, depth data and skeleton information: section III contains motivation of the proposed work: section IV describes our suggested method for activity recognition: experimental result and discussion is given in the section V and section VI, respectively, and at last conclusion and future work has been described in section VII.

IV. Proposed Approach

To represent human single-limb and multi-limb activities sequence recognition, this paper proposes a novel framework based on the color information retrieved from a RGB camera in an intelligent video surveillance system. We have done our proposed work in two stages. In the first stage input data have been preprocessed by resizing into the new scale, then we have extracted the histogram of oriented gradients (HOG) features from all input frames and we have classified the activities into two categories named single-limb and multi-limb activities, using a random forest (RF) classifier. Then in the second stage two 2D convolution neural networks have been applied to each category for recognition of actual activity type under single and multi-limb category. The flow diagram of the proposed two stage human activity recognition is depicted in Fig. 2.

A. Preprocessing and Feature Extraction

Resizing: Collected RGB frames from MSR UTKinect Action 3D dataset have size of 480x640. The proposed approach read all single and multi-limb activity sequences for resizing the initial raw color information to a new size (80x120) using the OpenCV library.

Histogram of Oriented Gradients: HOG features are widely used in various motion based applications and activity recognition systems. These features gradients have been calculated over preprocessed image. We have divided the image into 8x8 cells. All the gradient values of a cell are divided into 9 equal bins histogram. To estimate the feature value, a block size of [16, 16] has been taken. A [16, 16] block has 4 histogram which will form a one dimensional vector of size 36. In addition to this, a detection window size of [16, 16] is used having a stride of [8, 8] leading to the14 horizontal and 9 vertical positions where the detection window moves for constructing a total of 126 positions. To calculate the final feature vector of an entire image, all 126 features having a size of 36, are combined together to form a final large vector, which is a 4536(126 * 36) dimensional vector. The final feature vector (F) can be defined using (1) where d_1, d_2, d_3…… d_k are the dimensions.

$$F = (d_1, d_2, d_3, \ldots\ldots\ldots\ldots d_k) \qquad \text{where } k = 4536 \tag{1}$$

B. Initial Classification using Random Forest

Random Forest is a type of ensemble based classifier proposed by Breiman [1] in 2001. The RF method works on different models increasing the accuracy (bagging) and improving the performance of previous trees by the subsequent trees (boosting). The basic principle of Random Forest is that it takes the decision based on de-correlated decision trees. It can be used with multi-class classification purposes. An RF model is non-parametric in nature. For an ensemble of classifiers C1(θ), C2(θ),.....Cn(θ), and having a training set chosen arbitrarily from the distribution of the random vector P, Q, then the margin function given in equation (2) can be defined as,

$$mg(\mathbf{P}, Q) = av_n I(C_n(\mathbf{P}) = Q) - \max_{i \neq Q} av_n I(C_n(\mathbf{P}) = i) \tag{2}$$

Where $i \neq Q$

Where I (*) is the indication function. The margin determines the limit to which the average number of votes at **P**, Q for the right class exceeds the average vote for any other class. In our work, we have used the RF classifier to distinguish two classes, i.e. either single-limb or multi-limb activity. Since Random Forest is an ensemble classifier, a HOG feature has been computed for each video frame in an activity. This final feature vector (F) of dimension 4536 has been used to train the classifier.

C. 2D-CNN Classifier Based Activity Recognition

A 2D-CNN or ConvNet is a type of deep neural network, often used for video analysis and recognition and image classification. ConvNet is a biological inspired deep-network whose connectivity designs between the neurons are similar to the animal visual cortex.

CNN is a sequence modeling classifier and has a sequence of layers where every layer transforms input volume of activation into another form. In ConvNet, three main layers are used to make a ConvNet model, a convolution layer, pooling or sub sampling and fully connected layer. We keep the layers one after another to build the ConvNet architecture. CNN uses minimal preprocessing compared to other classification algorithms. Automatic feature detection is the considerable advantage. In this work, we have used 2D-ConvNet classifiers for recognition of activities that belong to the single limb and multi-limb classes.

Therefore, two 2DconvNet models have been trained using the categorical cross entropy (CCE) based objective function. The architecture of 2D-ConvNets for single and multi-limb activity is given in Fig. 3.

CCE based 2DconvNet: The Network has C output values, corresponding to one value of each class for the activity sequences. The Categorical Cross entropy (CEE) for C classes is defined in (4)

$$CCE = -\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{C} l_{yi \in Ck} \log Prob_{model}[yi \in Ck] \tag{4}$$

Where N represents the number of samples. Each sample belongs to a category k (the total of categories is C). $l_{yi \in Ck}$ *is the indicator function of j samples for belonging to* k^{th} category and logProb model $[y_j \in C_k]$ is the predicted-probability of the jth sample belonging to k^{th} category.

V. Experimental Work

First, the description of the dataset used in this study is given. Then results have been evaluated by splitting the dataset into a 80% training set and a 20% test set using the train test split method. For training and validation of our framework, we used the public dataset: UTKinect-Action Dataset. The model was trained on Intel Core i5 7th Gen, 2.4GHz processor, 8GB of RAM and a 2GB 940MX NVIDIA GPU support on Ubuntu 16.04 LTS(Linux) operating system.

A. UTKinect-Action3D Dataset

To validate our proposed framework we composed a dataset having 10 types of indoor human activity sequences. The dataset was taken using a stationary Kinect sensor with Kinect for Windows SDK Beta version having a frame rate of 30 fps. The Kinect sensor has a capacity to capture about 4 to 11 feet. The Dataset contains 10 activities performed by 10 different persons two times: 9 male and 1 female, out of them one person is left-handed and the rest are right-handed, with a total of 199 activity sequences. The label of the carry activity performed by the persons the 2nd time is not given, hence frames for this activity cannot be identified. The 10 activity sequences are wave hands, pull, walk, sit down, push, stand up, throw, pick up, carry, and clap hands. All these activity sequences are given in the three different formats RGB, Depth frames and skeleton information. The dataset contains all the actions in indoor scenario.

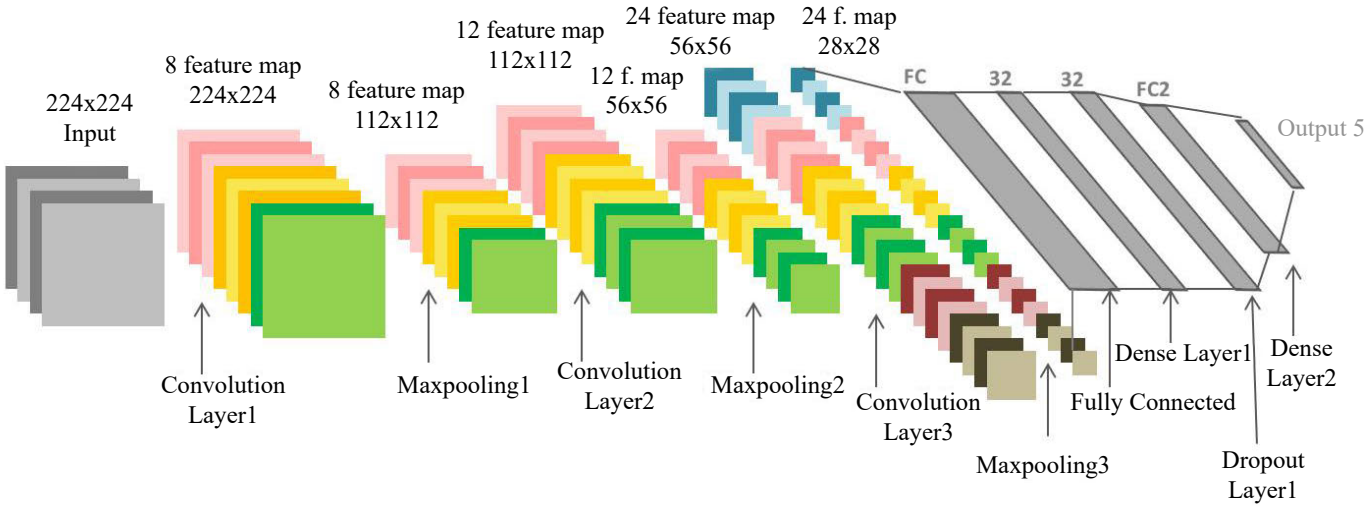
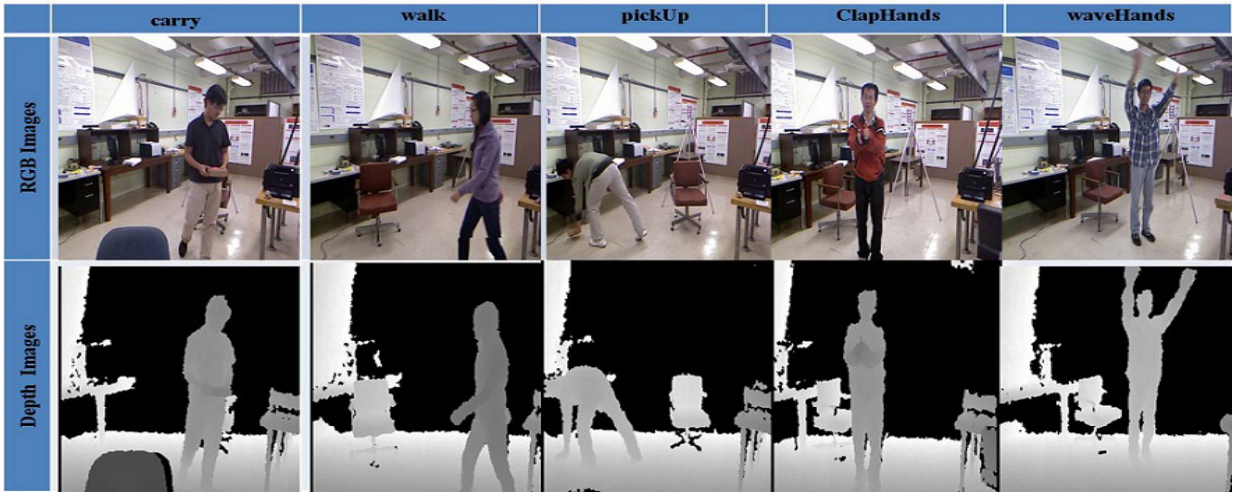
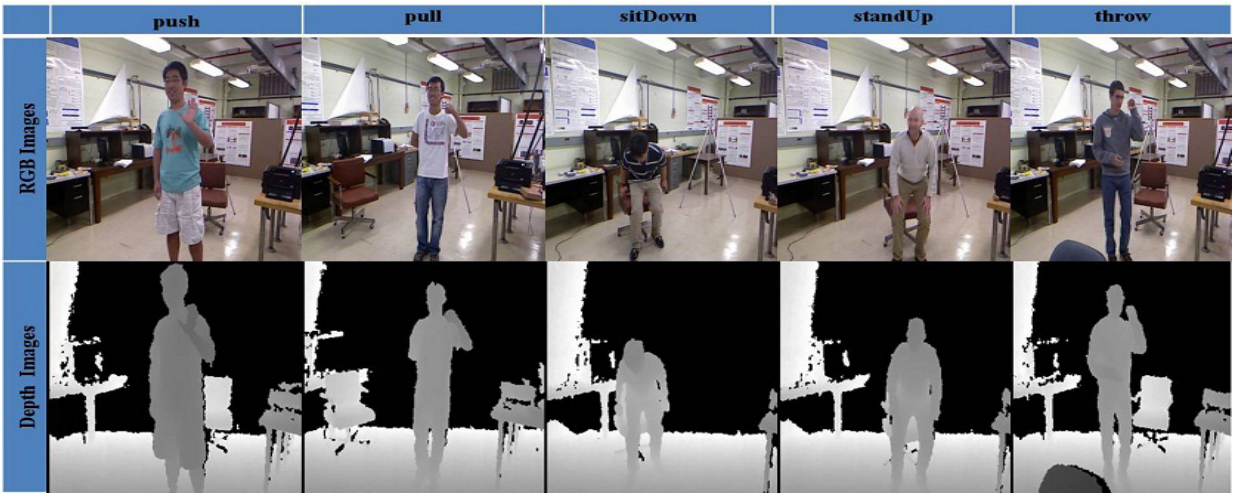


Fig. 3. 2D-ConvNet Architecture for human single and multi-limb activity recognition.



(a)



(b)

Fig. 4. Few sample frames of 10 different activity sequences are shown from the dataset. (a) Shows multi limb activities in RGB frames and their corresponding depth frames. (b) Shows single limb activities RGB frames and their corresponding depth frames. In this work only RGB frames are used for action recognition and depth frames are just for depiction.

The number of frames for each activity ranges from 5 to 120. The resolution of RGB and Depth images is 480x640. The total numbers of frames are 5869 for 199 activity sequences. The proposed frame work uses activity recognition only in RGB frames sequences and the rest of the sequences are just for demonstration of the dataset. Few RGB frames and their corresponding depth images are given in the Fig. 4.

B. Activity-Type Recognition using Random Forest

To Recognize single limb and multi limb types of activities we trained the random forest classifier from Scikit-Learn Python Library. The classification has been performed by varying the number of decision trees (n_estimators) from 1 to 100.An accuracy of 99.92% has been recorded in classification of single limb and multi-limb activity types for n=46.This is shown in the confusion matrix given in Fig. 5, that all samples are classified except one sample from single limb class.

C. Activity Recognition using 2D-ConvNet

Two 2D-ConvNet classifiers have been trained for single limb and multi-limb activities using the output of initial Random Forest Classification with the help of a feature vector F. For Single limb activity, the network has been trained with categorical cross-entropy objective function with a learning rate of 10^{-3} and decay of $5x10^{-6}$. The architecture of 2D-ConvNet is given in Fig. 3 and the learning curve of the network for single-limb activities is shown in the Fig. 6(a) and, for multi-limb activities the learning curve is given in Fig. 6(b).

It can be seen from the learning curve in Fig. 6(a) of single limb activity that, after 146 epochs, there is no change in the validation network. Thus, it has been pointed as the Best-Network. An Accuracy of 97.9% has been recorded in the recognition of single-limb activities as shown in Fig. 7(a). The Confusion matrix corresponding to classification of single limb activities is given in Fig. 8(a) Recognition performance has also been marked against each class of activities as shown in Fig. 9(a) where accuracies vary from 89% to 100% for different activities and 100 percent accuracies have been recorded for pull, stand up and throw activities. It also has been noticed that one more activity sit down is having approximately 100 percent recognition.

The learning curve corresponding to the multi-limb activities is shown in the Fig. 6(b). After 42 epochs, the best network is found because further there is no change in the validation results. An accuracy of 98% has been recorded in recognition of multi-limb activities given in Fig 7(b). The confusion matrix of multi-limb activities is given in Fig. 8(b). Recognition performance has also been recorded for every individual class of activities as depicted in Fig. 9(b). It is also noticed that 100 percent recognition is achieved for clap hands activity.

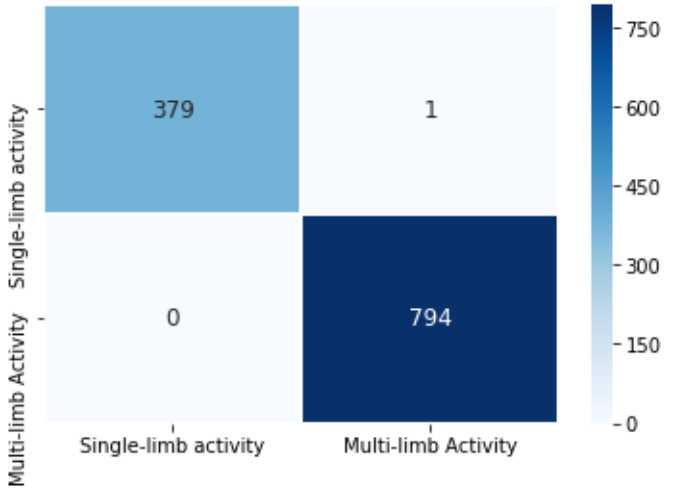
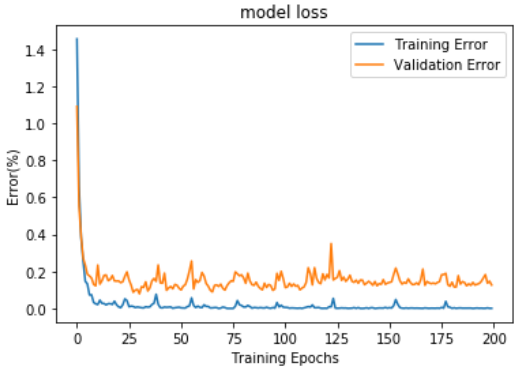
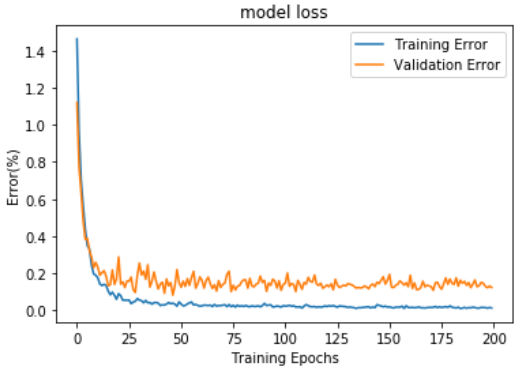


Fig. 5. Confusion Matrix of activity-type recognition.

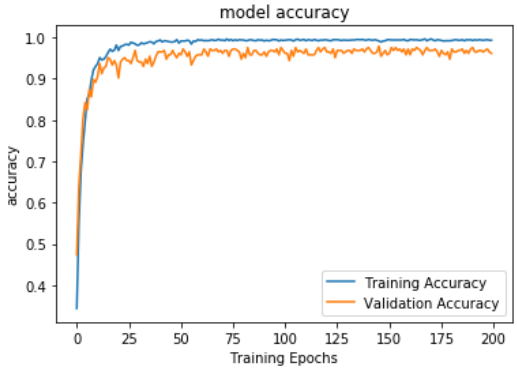


(a)

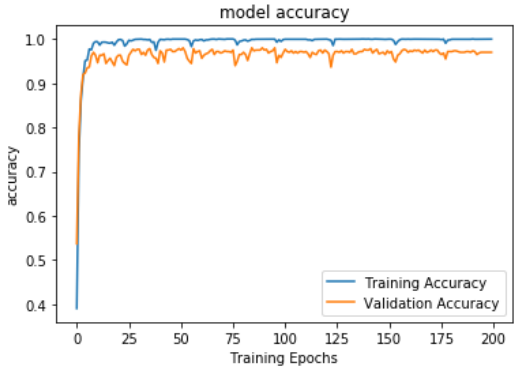


(b)

Fig. 6. Learning curves of 2DConvNet showing variation in training and validation (a) For single-limb activity (b) For multi-limb activity.

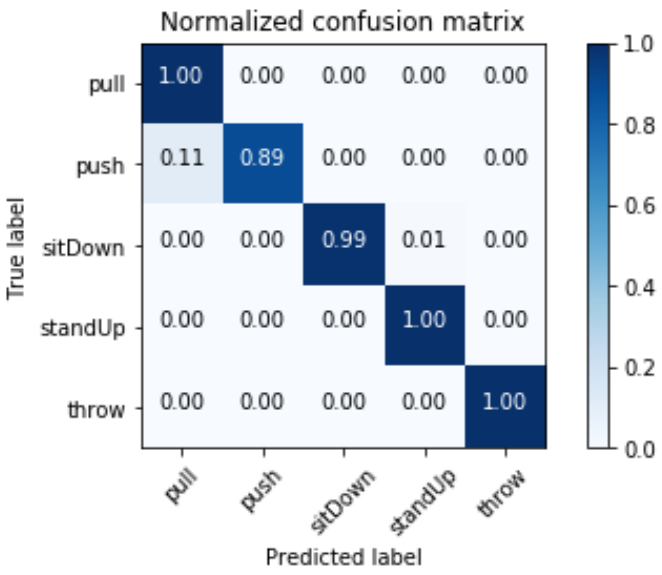


(a)

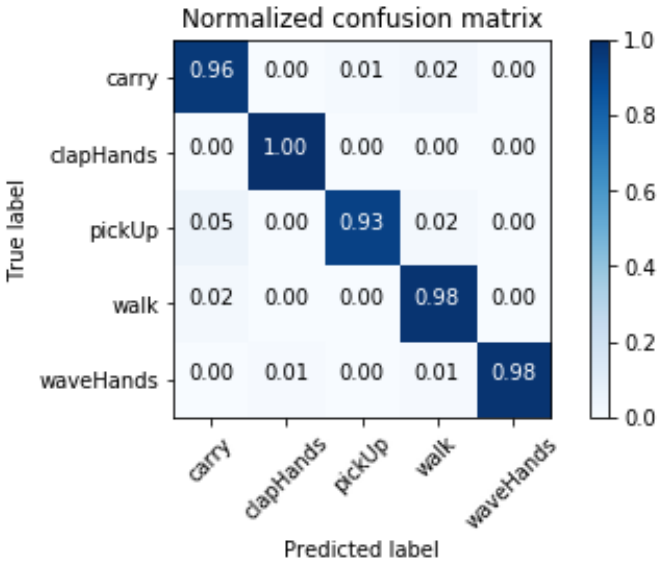


(b)

Fig. 7. Accuracy curves of 2DConvNet (a) shows accuracy curve of single-limb activity (b) shows accuracy curve of multi-limb activity.

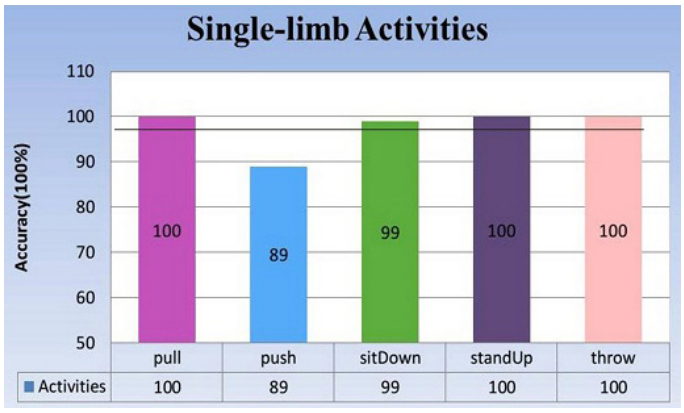


(a)

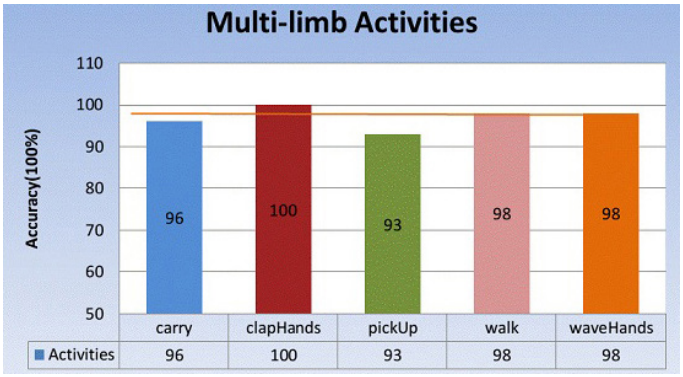


(b)

Fig. 8. Confusion matrix of 2DConvNet model (a) For single-limb activities (b) Multi-limb activities.



(a)



(b)

Fig. 9. Activity recognition performance for each activity class: (a) single limb activities (b) multi limb activities.

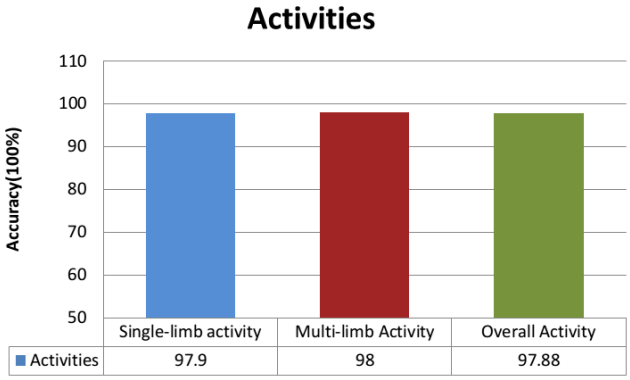


Fig. 10. Comparative performance analysis between recognition rates of single-limb and multi-limb human activities along with complete system performance..

VI. Results and Discussions

This section presents the obtained experimental results and discussions about them. In this paper we adopted two-stage classification in which we obtained the results in two phase called coarse and fine level classification. First stage classification divides the activity types in two categories, single-limb and multi-limb respectively. Second stage classification has been performed to recognize actual type of activities in both. Finally a randomly train test split method has been used to validate our proposed approach.

A. First Stage Classification Results

In the first level classification, human activity types are categorized into two type's, single-limb and multi-limb activities. This classification makes easy the process of recognition at the second level. In this phase classification has been performed by Random Forest classifier changing the number-of-trees in the forest. It is illustrated from the confusion matrix given in Fig. 5 that all test samples have been classified correctly into single-limb and multi-limb classes except for one sample and 99.92% is the classification accuracy.

B. Second Stage Classification Results

In the second stage, actual recognition of individual activities takes place. Recognition has been performed using human activities for both single-limb and multi-limb categories. The overall recognition accuracy of single-limb activities is 97.9%. Individual activity classes named throw, standup and pull have been recognized 100 percent correctly, and for the remaining two activities push and sitdown the obtained accuracy is 89% and 99 %, respectively. On the other hand the

overall recognition accuracy of multi-limb activities is 98% in which claphands activity is recognized 100% and the remaining individual classes named carry, pickup, walk and wavehands have 96%, 93%, 98% and 98% accuracies, respectively. It has also been illustrated from Fig. 10 that the overall accuracy of the system is 97.88 %, which is better than some of the previous state-of-art results.

The approach proposed in the paper is better than the existing methods in the sense that in a two-stage strategy the larger class problem may be divided into the sub class problems where individual sub classes are recognized at next sub level in downward stream, since real time recognition is difficult if larger number of classes are taken into account. Hence, the problem of large classes may be improved by dividing the classes into two or more levels. The recognition performance of multi-level classification depends on the training losses incurred at each level. Therefore we try to minimize the training losses at each level of classification. For example the author in [49] performed classification task on EEG signals dataset by considering the whole dataset at a time and got 39.34% accuracy, but when they used two-stage strategy (coarse-fine level classification) they got 85.20 % accuracy at coarse level and 65.03 % at fine level and combined accuracy was 57.11% (85.20 % * 65.03 %), a major increment in accuracy by a factor of 17.77% using two-stage classification.

C. Error Analysis

The proposed human activity recognition system is working in two stages. In the first stage, the proposed system is classifying single limb and multi limb activities with an accuracy of 99.92% which is approximately 100% except only one sample whose actual class was single-limb, which was being predicted as multi-limb.

In the next stage, after classifying single and multi-limb activities when each individual category is being recognized, 11% of the push activity in single limb was erroneously recognized as pull, owing to the similarity in both activities because the captured video frames show similarity during these two activates. Simultaneously, 1% error was detected (during sit down activity) as standup activity due to the high degree of similarity between the two activities.

While in case of multi-limb activity, an error of 4% was found in carry activity, which is misclassified 1% as pick up and 2% as walk, because carry activity is somewhat a combination of walk and pick up activities. Similarly, 7% of misclassification error was found in the pickup activity, being 5% of instances predicated as carry and 2% as walk, because in pickup activity video frames the subject is carrying some goods and some pickup activity video frames are similar to walk activity. At the same time, an error of 2% was found in walk activity and model predicted it as a carry activity because both activities contain the motion information. Similarly 2% misclassification was in wave hands activity which is predicted 1% as wave hands activity and 1% as walk activity due to the similarity between them.

D. Comparison with State-of-art

We compared classification-accuracy of our proposed system with other approaches given in previous methodologies. The result comparison has been shown in Table II. Our method achieved the highest accuracy among the methods given in Table II. Starting from [29] where the authors have taken human posture as histogram of 3D joints (HOJ3D) as a novel descriptor and got 90.92% classification accuracy while our two-stage strategy produces 97.98 % on the same dataset. Liu et al. [46] used both RGB and depth information of human activities and fussed this information together with coupled hidden conditional random field model and generated 92% accuracy. Zhao. et al. [47] used raw depth sequences, depth motion map and RGB information and fussed together all this information and applied 3DSTCNN with SVM for human action recognition. The proposed approach in [48] produces 97.29% accuracy on UTKinect-Action 3D dataset, 94.15% accuracy on MSR-Action 3D dataset. Vemulapalli. et al. [48] used 3D geometry of different body parts using translation and rotation in 3D space and generated 97.08% accuracy on the UTKinect-Action 3D dataset. It is clear that our proposed approach generates good results on the UTKinect-Action 3D dataset as compared to the methods given in Table II. Thus our methodology advances some of the methodology as discussed above.

TABLE II. Performance Comparison with Other Methods on UTKinect-Action3D Dataset

Methods	Accuracy
Xia. et al., (2012) [29]	90.92 %
Liu et al., (2015) [46]	92.00 %
Zhao. et al., (2019)[47]	97.29 %
Vemulapalli. et al., (2014) [48]	97.1 %
Proposed Approach	**97.88%**

Based on the comparison with previous state-of-art results discussed in Table II our proposed approach has some advantages which are mentioned below:

- The multi-stage method facilitates the classification task by reducing large training losses given in complex problems into the low training losses given at the different levels.
- Complexity of the system may be reduced by making multi stages.
- Better recognition using initial and subsequent levels.
- Suitable for human computer interaction application.

Although our proposed multi-stage strategy has good results as compared to the state-of-art results given in Table II, it also has some limitation as the system will produce good results if and only if the classification accuracy at the initial stages is high.

VII. Conclusion

This paper presents a novel framework to recognize human single-limb and multi-limb activities using video frames. This framework facilitates to analyze human limb activities in real-time. The recognition process has been done in two stages. Firstly a Random Forest classifier has been used to distinguish input activities into two classes of activities, such as human single-limb and multi-limb. In the second phase, two 2D Convolution neural network classifiers have been trained for recognition of separated activities using a sequence classification based approach. The UTKinect-Action Dataset of 199 activities sequences has been used by the proposed framework. An accuracy of 99.92% was achieved using the Random Forest classifier. An overall accuracy of 97.88% has been recorded by our system for both types of activity classes. The major components of this proposed approach are real time, computation of HOG feature and classification. Obtained experimental results show the major advantage of deep convolution neural network implementation in activities recognition. This work also proposes the advantages of applying RGB information to recognize human activity types. In future work, Depth frames and Skeleton joints data may be combined with RGB information to form a large amount of data and generate a robust approach for better human activity recognition.

Acknowledgment

We are thankful to Uttarakhand Technical University Dehradun for maintaining the research facilities for this work. We also give special thanks to Microsoft Research team to maintain UTKinect-Action Dataset repository to complete this research work. We also thanks to

Dr. Pradeep Kumar Postdoctoral Fellow at UNB Canada for revising and improving this manuscript

References

[1] L. Breiman, "Random forests". Machine Learning, vol. 45, Jan. 2001 pp. 5-32.

[2] N. Haering, P.L. Venetianer and A. Lipton, "The evolution of video surveillance: an overview", Machine Vision and Applications, vol. 19 May 2008, pp. 279-290.

[3] W. Hu, T. Tan, L. Wang and S. Maybank, "A survey on visual surveillance of object motion and behaviors", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol 34, 2004, pp. 334-352.

[4] I.S. Kim, H.S. Choi, K.M. Yi, J.Y. Choi and S.G. Kong, "Intelligent visual surveillance a survey", International Journal of Control, Automation, and Systems, vol. 8, 2010, pp. 1598- 6446.

[5] T.Ko, "A survey on behavior analysis in video surveillance for homeland security applications", in: 37th IEEE Applied Imagery Pattern Recognition Workshop, 2008 (AIPR 08), October 2008, pp. 1-8.

[6] O.P. Popoola and K. Wang, "Video-based abnormal human behavior recognition review", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 42, 2012, pp. 865-878.

[7] K. K. Verma, P. Kumar and A. Tomar, "Analysis of moving object detection and tracking in video surveillance system", In 2015 IEEE 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1758-1762.

[8] P. Geetha, V. Narayanan, "A survey of content-based video retrieval", Journal of Computer Science, vol. 4 2008, pp. 474-486.

[9] J. Zhang, Z. Liu, (2008, June). "Detecting abnormal motion of pedestrian in video", In IEEE International Conference in Information and Automation ICIA, 2008, pp.81-85.

[10] C. C. Hsieh, and S. S. Hsu, "A simple and fast surveillance system for human tracking and behavior analysis." In Third IEEE International Conference on Signal-Image Technologies and Internet Based System SITIS'07, December 2007, pp. 812-818.

[11] J.K. Aggarwal and Q. Cai, "Human motion analysis: a review", Computer Vision and Image Understanding, vol. 73, 1999, pp. 428-440.

[12] J.K. Aggarwal, Q. Cai, W. Liao and B. Sabata, "Non rigid motion analysis: articulated and elastic motion", Computer Vision and Image Understanding, vol. 70, 1998, pp. 142-156.

[13] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: a review", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, 2010, pp. 13-24.

[14] R. Poppe, "Vision-based human motion analysis: an overview", Computer Vision and Image Understanding, vol. 108, 2007, pp. 4-18.

[15] W. Wei and A. Yunxiao, "Vision-based human motion recognition: a survey", in: Second International Conference on Intelligent Networks and Intelligent Systems, 2009 (ICINIS09), November 2009, pp. 386-389.

[16] H. Buxton, "Learning and understanding dynamic scene activity: a review", Image and Vision Computing, vol. 21, 2003, pp. 125-136.

[17] M. Del Rose and C. Wagner, "Survey on classifying human actions through visual sensors", Artificial Intelligence Review, vol. 37, 2012, pp. 301-311.

[18] M. Pantic, A. Pentland, A. Nijholt and T. Huang, "Human computing and machine understanding of human behavior: a survey", in: Proceedings of the 8th International Conference on Multimodal interfaces, ICMI 06, ACM, New York, NY, USA, 2006, ISBN 1-59593-541-X, pp. 239248.

[19] W. McNally, A. Wong, and J. McPhee, "STAR-Net: Action Recognition using Spatio-Temporal Activation Reprojection", 16th IEEE Conference on Computer and Robot Vision (CRV), 2019, pp. 49-56.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-scale video classification with convolutional neural networks". In Proc. CVPR, 2014, pp. 1725-1732.

[21] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. "Convolutional learning of spatio-temporal features". In Proc.ECCV, 2010, pp. 140-153.

[22] Tran, L. Bourdev, R. Fergus, L. Torresani, and M.Paluri. "Learning spatiotemporal features with 3D convolutional networks". In Proc. IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489-4497.

[23] L. Xue, S. Xiandong, N. Lanshun, L. Jiazhen, D. Renjie, Z. Dechen, and C. Dianhui, "Understanding and Improving Deep Neural Network for Activity Recognition" .arXiv preprint arXiv: 1805.07020, 2018.

[24] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: "A unified embedding for face recognition and clustering". In Proc. IEEE International Conference on Computer Vision and Pattern Recognition CVPR, 2015, pp. 815-823.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". In Advances in Neural Information Processing System (NIPS), 2012, pp. 1097-1105.

[26] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos". In Advances in Neural Information Processing System (NIPS), 2014, pp. 568-576.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions". In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9.

[28] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks", In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 648-656.

[29] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints". In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, June 2012, pp.20-27.

[30] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks". IEEE access, vol. 6, pp. 17913-17922.

[31] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A.Oliva, "Learning deep features for scene recognition using places database", in Proc. Advances in Neural Information Processing System, 2014, pp. 487-495.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2014, pp. 580-587.

[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deep Face: Closing the gap to human-level performance in face verification", in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701-1708.

[34] J. Donahue, A. Hendricks, L. Guadarrama, S. Rohrbach, M. Venugopalan, S. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description", in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625-2634.

[35] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition", IEEE Transaction on Pattern Analysis Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan.2010.

[36] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors", in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305-4314.

[37] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets, CoRR", In IEEE International Conference on Computer Vision and Pattern Recognition, July 2015, pp. 1-5.

[38] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition", in Proc. IEEE Conference on Computer Vision and Pattern Recognition., 2016, pp.3034-3042.

[39] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition", in Proc. IEEE Conference on Computer Vision and Pattern Recognition., 2016, pp. 1933-1941.

[40] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice", Computer Vision and Image Understanding, vol. 150, 2016, pp. 109-125.

[41] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors". In Proc. European Conference on Computer Vision (ECCV), 2014, pages 581-595.

[42] H. Wang and C. Schmid, "Action recognition with improved trajectories, In Proc. International Conference on Computer Vision (ICCV), 2013, pp. 3551-3558.

[43] H. Wang, A. Klser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories", In Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3169-3176.

[44] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance", In Proc. European Conference on

Computer Vision (ECCV), 2006, pp. 428-441.
[45] L. Sun, K. Jia, D.-Y. Yeung, and B. Shi, "Human action recognition using factorized spatio-temporal convolutional networks", In Proc. International Conference on Computer Vision (ICCV), 2015, pp. 4597-4605.
[46] A. A. Liu, W. Z. Nie, T. Su, L. Ma, T. Hao, and Z. Yang, "Coupled hidden conditional random fields for RGB-D human action recognition." Signal Processing, vol. 112, 2015, pp. 74-82.
[47] C. Zhao, M. Chen, J. Zhao, Q. Wang, and Y. Shen, "3D Behavior Recognition Based on Multi-Modal Deep Space-Time Learning." Applied Sciences, vol. 9.no. 4, 2019, pp. 7-16.
[48] Vemulapalli, Raviteja, Felipe Arrate, and Rama Chellappa. "Human action recognition by representing 3d skeletons as points in a lie group." Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 588-595.
[49] P. Kumar, R. Saini, P. P. Roy, P. K. Sahu, and D. P. Dogra, "Envisioned speech recognition using EEG sensors." Personal and Ubiquitous Computing, 2018, vol. 22, no. 1, pp. 185-199.
[50] P. Sewaiwar and Kamal Kant Verma. "Comparative study of various decision tree classification algorithm using WEKA." International Journal of Emerging Research in Management & Technology, vol. 4 2015, pp. 2278-9359.
[51] Siirtola, Pekka, and Juha Röning. "Revisiting" Recognizing Human Activities User-Independently on Smartphones Based on Accelerometer Data"-What Has Happened Since 2012?" International Journal of Interactive Multimedia & Artificial Intelligence, 2018, vol. 5, no. 3, pp. 17-21.
[52] A. Jalal and S. Kamal. "Improved behavior monitoring and classification using cues parameters extraction from camera array images." International Journal of Interactive multimedia and Artificial Intelligence, 2018, vol. 5, no. 3, pp. 2-18.

Kamal Kant Verma

Kamal Kant Verma is a research scholar in Uttarakhand Technical University Dehradun. He is currently working as an Assistant Professor in the Department of CSE, College of Engineering Roorkee Roorkee Uttarakhand India. He has 13 years of Teaching and research experience. His research area is Computer Vision, Video Processing, Machine Learning, and Deep Learning.

Brij Mohan Singh

Brij Mohan Singh is Dean Academics & Professor in Department of CSE, COER Roorkee. 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 Digital Image Processing and Pattern Recognition. He has guided 3 PhD Thesis of Uttarakhand Technical University (UTU) Dehradun India and currently 6 are in process.

H. L. Mandoria

H. L. Mandoria is Professor & Head in Department of IT, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar. He has published more than 70 research papers in International Journals and conferences. His research area is Computer Network, Information Security and Cyber Security. He has guided more than 20 M.Tech and presently three phd are in progress.

Prachi Chauhan

Prachi Chauhan is a research scholar in Govind Ballabh Pant University of Agriculture and Technology, Pantnagar. Her research area is Cyber Security, Machine Learning, and Deep Learning.