IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Special Issue-7, Sept. 2018 E-ISSN: 2347-2693

Human Activity Recognition using Deep with Gradient Fused Handcrafted Features and categorization based on Machine Learning Technique

G. Augusta Kani[1*], P. Geetha[2], A. Gomathi[3]

[1,2,3]Department of Information Science and Technology, Anna University, Chennai, India

*Corresponding Author: augus.jesus@gmail.com

Available online at: www.ijcseonline.org

Abstract – Human action recognition (HAR) from videos is a significant and has more research focus in the domain of Computer vision. The purpose of human action recognition in videos is to detect and recognize the human actions from the sequence of frames. Human action recognition undertakes many difficulties such as differences in human shape, cluttered background, moving cameras, illumination conditions, motion, occlusion, and viewpoint variations. In previously, local features or deep learned features are used to recognize the action. In the proposed work, both the features are used to recognize action and for analysis. From sequences of frames background is subtracted using Multi-frame averaging method. Two kinds of feature extraction are done. Shape based feature extraction, Optical flow feature extraction are some of the hand-crafted features performed and classification is done using HMM. The other one is deep learned features. Convolutional Neural Network extracts the features from frames in each layer. It extracts the features such as line, edge, color, texture and Classification is done using SVM. For human action recognition, hand-crafted features attain good result but it fails on large set of data. Deep learned features such as CNN have been used for large dataset and good result is obtained on recognition. To improve the human action recognition result, CNN is proposed. We compared both the approaches CNN and HMM and the results were analyzed. CNN results better accuracy while comparing with HMM.

Keywords – Background Subtraction, Convolutional Neural Network, Canny Edge Detection, Optical Flow, Hidden Markov Model

I. INTRODUCTION

In Computer Vision HAR plays an essential role in research area. Computer vision can be defined as “the concept and skill for building artificial systems that obtain info from images”. Computer vision is a field that includes approaches for finding, processing, examining images and that extract information from images and videos. Human actions are composed of complex actions, gestures and the actions also differ into spatial and temporal information. Background subtraction is the method of extracting the region of interest from the background for the sequence of frames.

The video features of human action recognition are classified into hand crafted features and the deep learned features. The Convolutional Neural Network(CNN) is mainly used for image classification problems, it is a deep learned feature. The number of layers included is depended on the particular application. Convolutional Neural Network extracts the features from each layer. There will be multiple hidden layers and a normalized output layer.

The hidden layer transfer the input several times and normalize the output layer. The Gradient descent algorithm is used to find the highest accuracy and minimize the prediction error. In HAR, usually they will represent the action, detect and segment the human action.

The frame itself saves sufficient information for human action recognition. The CNN requires large amount of training data to avoid the overfitting. CNN extract sub image activations and some CNN layer feature activations. The activation layer control how the motion flow from one layer to next layer. SVM classification is used for classifying the frames and finding the accuracy by training and validating the frames.

Activity recognition can be achieved by differentiating the activities based on its appearance, shape, interest point, optical flow or motion features. Human action changes dynamically. Temporal templates representation captures both motion and shape features to recognize human action.
Shape can be represented by the border, area, moment, etc. Canny edge detection technique is used to detect the edges of the image. Optical flow is the form of seeming motion of objects, surfaces and edges from the sequence of frames caused by the relative motion between the current frame and the next frame. For classification, Hidden Markov

Model(HMM) is used. A HMM consists of a amount of states each of which is allocated a possibility of transition from one state to another state. HMM classifier was trained for different activities using Baum Welch algorithm and for testing maximum likelihood estimation is used for classifying the activity.

The paper is organized as follows. Section II summarizes the work related to the paper. Section III explains proposed methods used for HAR. Section IV explains about the techniques used and Section V discusses results and analysis of this proposed work.

II. RELATED WORK

Trajectory-Pooled Deep-Convolutional Descriptors technique have been proposed by [4], that shares the advantage of both hand-crafted features and deep learned features. In the hand-crafted features, the improved trajectories combined with Fisher vector are most successful. In deep-learned features, Convolutional Networks, which contains a sequence of convolutional and pooling layers and CNN automatically learn features with a trained neural network. Bag of Visual Words model (BoVW) with local features are used [5]. BoVW constructed a pipeline for converting a set of local features into global representation. The two-stream network is mainly used to speed up the network as it reduces the input dimensionality in half. The paper [6] proposes 4 different types of fusion, namely, Single-frame, Early fusion, Late fusion, Slow fusion. The paper [8] proposes the long-term recurrent convolutional networks (LRCN) architecture to learn video representation that can be used in various tasks. The LRCN model is a combination of convolutional neural networks and recurrent neural networks. Having a RNN permits the network to accept a sequence as its input. The CNN produces a representation of an inert frame that is combined with long term temporal information carried by a sequence of frames. Stratified Pooling method have been used [7] it transforms the random number of frame-level features to fixed length of video-level features. The video level features are done by averaging the sampled video in 3 levels. In each level the sample video volume is divided into sub-volumes and the 3 levels are concatenated into video level features. The hand-crafted feature has many useful properties that are used in CNN [9]. The hand-crafted features are composed of two stages the convolution and the pooling layer in CNN. The deep learned features are used for solving complex learning problems in various domain. Multiclass SVM Classification [7] have been used and it achieves the highest accuracy on the two datasets HMDB-51 and UCF-101 for human action recognition by multiclass combination. The Support Vector Machine (SVM) is a powerful discriminant classifier first developed in paper [10]. CNNs were found to be excellent feature extractors for other classifiers such as SVMs [11]. This generally involves taking the output of the lower layers of the CNN as feature extractors for the classifier. A two-layered background subtraction [1] is used, which is based on both chromaticity and gradient to extract human contours from the frame sequence captured by camera. This subtraction helps us to remove shadows from foreground and to get a good contour for recognition. In the paper [2] background subtraction is performed using Gaussian mixture model. Kalman filter which tracks the evolution of a single Gaussian, the mixture of Gaussian methods tracks multiple Gaussian distributions. Mixture of Gaussian maintains a density function for each pixel. Adaptive background subtraction algorithm has been implemented to obtain global outline feature and optical flow model to extract local visual feature [3]. Then these feature vectors are combined to form a hybrid feature vectors. The shape descriptor is used to describe the image content [12] has been mainly used canny edge detection for shape based feature extraction. An action recognition method has been proposed using Symbolic Aggregate approximation(SAX) [13]. A frame was transformed into time series and these time series where converted into a SAX vector for action recognition. Optical flow algorithm such as the Lucas Kanade approach more robustness to noise than dense optical flow algorithms. A sparse optical flow algorithms [15] has been implemented to evaluate the movement for a certain number of pixels in the image. The Lucas Kanade technique is a commonly used variance way that has been implemented for optical flow estimation [14]. This method assumes that the flow is persistent in a local neighborhood of the pixel. This method is less sensitive to image noise than point-wise methods. Labeled the activities by means of a Compound Hidden Markov Model [18]. Each distinct Linear Hidden Markov Model has wave information of a human activity. The sequence of maximum possible states from a sequence of observations, indicates which activities are done by a person in a interval of time. An action recognition method based on HMM using the Bag of Words method. Time sequential images of human action are transformed into feature vectors. In The paper [12] has been stored all the feature vectors in a codebook where each symbol corresponds to an action by using vector quantization (VQ). From each pair of activity video images acquired by a stereo camera, the body joint angles are estimated by co-registering a 3-D body model to the stereo information. The paper [16] has been mapped the estimated angle features from the time sequential activity video frames into code words to generate a sequence of discrete symbols for a Hidden Markov Model of each activity.

III. PROPOSED APPROACH

The system architecture as shown in Figure 1 illustrates the overall process in Human Action Recognition from videos using Convolutional Neural Network and Hidden Markov Model. The videos are converted into sequence of frames. The Region of Interest (RoI) is extracted by subtracting the background using multi-frame averaging method. After preprocessing the features are extracted using CNN and the extracted features are classified using Multiclass SVM Classification. The other part is feature extraction using Shape based feature extraction and Optical Flow feature extraction and two features are fused together using gradient fusion function. Thegradient fusion is used for HMM classification.

A. Background Subtraction

Multi frame averaging method is used for background subtraction. The video is down experimented and averaged from starting to finish. A background model is constructed and that is used for subtracting each frame in the video to extract region of interest. The foreground object can be obtained by subtracting the background, using average method. The background subtraction has been useful for all types of circumstances with various improved changes. Continuous and effective informing of the background in response to gradual changes of background. The region of interest is the part of the image which catches the attention instantly than the other parts of the image. Background subtraction is done using equation (1) and (2) where x is the current frame and y is the next frames is subtracted with respect to time.

$$B(x,y,t) = I(x,y,t-1) \qquad (1)$$

$$|I(x,y,t)) - I(x,y,t-1)| > Th \quad (2)$$

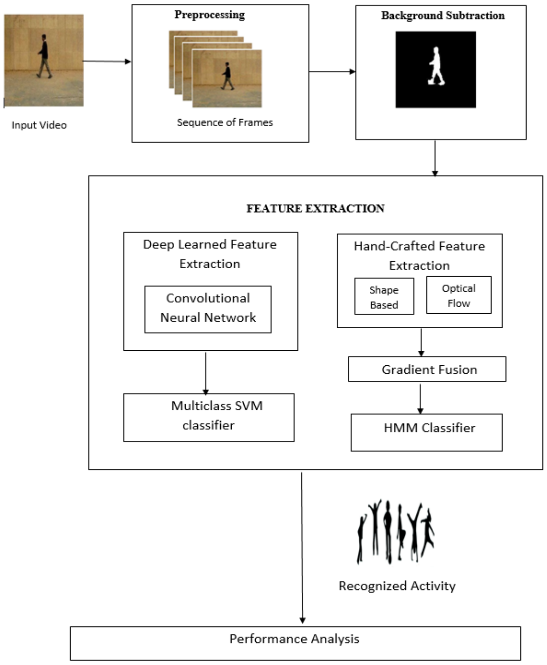

Figure 1: System Architecture for Human Action Recognition

B. Feature Extraction using CNN

The features such as HOG, SIFT, SURF, LBP are extracted for Human Action Recognition (HAR) Convolutional Neural Network (CNN) the features are extracted by specifying filter, stride, padding. In this work, the network is based on the AlexNet model. The AlexNet model consists of 25 layers. Mainly it as 5 convolutional layer, 3 fully connected layer and a SoftMax layer, other than it as ReLU (Rectified Linear Unit) layer, Normalization layer, Maximum pooling layer, Dropout layer, Classification layer.

CNNs eliminate the need of hand-crafted feature extraction; the features are extracted directly from sequence of images. CNN convolves learned features with input data.

a)Image Input Layer

Image input size defines the size of the input image of a CNN and contains the raw pixel value of the image. The size of an image corresponds to the height, width and the number of color channels of that image. Basically, the image size of the image is 320 x 240. At training time, the input of the network is fixed to 227 x 227 x 3.

b) Convolutional Layer

The first layer that receives an input signal is called convolutional filters. It is a process where the network tries to label the input signal. A convolutional layer consists of neuron that connect to sub regions of the input images. A convolutional layer learns the features localized by these regions while scanning through an image. The convolutional features are generic features, such as edge, color, texture features. A weight that are applied to an area in the image called a filter. The filter passes along the input image perpendicularly or parallelly. The extent with which it passes is called a stride. A filter transfers with the input, it uses the identical set of weights and bias for the convolutional, creating a feature map. The number of feature maps a convolutional layer has is equal to thenumber of filters. Each feature maps have a different set of weights and a bias. The convolutional layer output is calculated using equation (3) and when the padding size is 0. If the padding size is not zero equation (4) is used. Figure: 2 shows the feature extraction of the convolutional layer wherethe input size is 32, filter size is 5 and stride is 1.

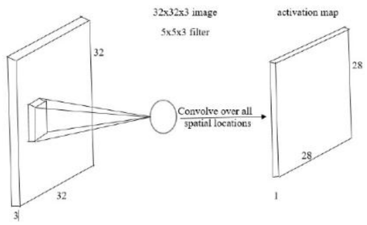

Figure 2: Convolutional Layer

$$Convolutional\ Layer = \left(\frac{Input\ size - Filter}{Stride}\right) + 1 \quad (3)$$

$$Convolutional\ Layer = \frac{Filter - Stride}{2} \quad (4)$$

c) Pooling Layer

The pooling layer consists of two types maximum pooling layer and average pooling layer. The pooling layer is used to control the overfitting. It reduces the number of parameters. It reduces the spatial size of the representation and it also reduces the amount of computation time. There are two types of pooling layer: maximum pooling layer and average pooling layer. The value is calculated using filter size 2 for the pooling layer using equation (5).

$$Pooling\ Layer = \left(\frac{Input\ size - Filter}{Stride}\right) + 1 (5)$$

1) Maximum Pooling Layer

Maximum pooling layer, returns the maximum value of the rectangular region of its input. In the feature map the maximum value is taken.

2) Average Pooling Layer

Average pooling layer, returns the average value of the rectangular region of its input. In the feature map the average value is taken.

d) ReLU Layer (Rectified Linear Unit)

A convolutional layer is generally seen as a non-linear activation function. It makes a threshold process to individual part. If any input value less than zero is set to zero and does not change the size of its input.

e) Dropout Layer

For the given probability, a dropout layer casually sets the layers input element to 0. For each new input element, the training network randomly selects a subset of neurons, forming a different layer of architecture.

f) Fully Connected Layer

The convolutional layers are followed by one or more fully connected layers. All neurons in a fully connected layer connects to the neurons in the layer previous to it. Fully connected layer combines all of the features learned by the previous layer across the image to identify the layer patterns. It is mainly used for classification; the last fully connected layer combines them to classify the image.

C) Shape based Feature Extraction

The shape method capture the local features from the human image. Shape can be signified by the border, area, moment, ends, etc., The feature extraction is based on edge point and texture. Canny edge detection is used for shape based feature extraction. The Canny edge detector is an edge identification operator that uses a multi-stage algorithm to detect a wide range of edges in images.

$$|G| = \sqrt{G_{x^2} + G_{y^2}} \qquad (6)$$

$$|G| = |Gx| + |Gy| \qquad (7)$$

By utilizing the law of Pythagoras as shown in Equation (6) which is a Euclidean distance, the edge strengths can be verified and equation (7) is Manhattan distance can be used to measure.

D)Optical Flow based Feature Extraction

An optical flow algorithm calculates the displacement of brightness patterns from one frame to another. This is done by intensity values of neighboring pixels. Optical flow is the seeming motion of intensity designs in the image. Lucas-Kanade method is used for optical flow algorithm. It estimates the displacement for a selected number of pixels in the image. Recover image motion at each pixel from spatial-temporal image brightness variations. It is used for extracting some special effects and estimating the dimensional, structure of the scene. The equation for the optical flow all the pixels in a positioned. The velocity vectorV_x and $V_y$of the local image flow must satisfy the equation (8).

$$I_x(q1) \times V_{x+} I_y(q1) \times V_y = -I_t(q1) \qquad (8)$$

Ix and *Iy*are partial derivative of image in x and y direction.

The equation (8) can also be written as:

$$P \times V = s \qquad (9)$$

E) Gradient Fusion

A hybrid approach combine the shape and optical flow based features together for human activity recognition. Shape and optical features were extracted from background subtracted image. The function Gradient Fusion have been used to fuse the frames. By combining both the feature extracted frames more information is obtained for classification.

This equation (10) used to compute gradient magnitude. It fuses the gradients of several single channel images by taking the maximum gradient magnitude at each pixel location.

$$(gxH^2 + gyH^2)\text{^}0.5 \qquad (10)$$

F) Multiclass SVM Classification

SVM for multiclass classification (one vs all classifier). The input belongs to one of the k classes. In one vs all, the training fits one classifier per class against all other data as a negative class in the total k classifiers. The prediction applies k classifiers to a new data point. In Cross Validation, the input is the images of three categories images and create a k-fold partition of the dataset. For each of k experiments, use k-1 folds for training and the remaining one for testing.

For each of the N classes the density is $P_k(x)$. The density is predicted using equation (11). In one vs all method, k indicates the labels, x indicates the samples.

$$y = arg\ \max_{k \in 1,\ldots N} P_k(x) \qquad (11)$$

Multiclass SVM-SGD (Stochastic Gradient Descent (SGD) algorithm uses the one-versus-all approach to train independently k binary classifiers. Two ways for creating the new multiclass SVM-SGD algorithm can handle large number of classes in high speed. The first one is to build balanced training of binary classifiers with a sampling approach. For one vs all approach equation (12) is used.

$$y = \arg\max_{k \in 1\ldots,k} f_k(\mathrm{k}) \qquad (12)$$

G) Hidden Markov Model

A HMM involves of an amount of states each of which is allocated a possibility of transition from one state to another state. With time, state transitions happen stochastically. Hidden Markov Model is used for classification. HMM is a stochastic state transit model which is robust against temporal, spatial and view-point variations. HMM classifier was trained for different activities using Baum-Welch algorithm and for testing maximum likelihood estimation is used for classifying the activity. Symbols are emitted from the states according to the assessed possibility assigned to them. HMM states are not directly identified only through a sequence of observed symbols. The maximum likelihood is used to discover the values that makes the identified numbers maximum possible.$a_{ij} = \frac{C(i\to j)}{\sum_{qeQ} C(i\to j)}$ (13)

The maximum likelihood estimation of the possibility a_{ij} of a specific transition among states i and j by calculating the number of times the transition was occupied and then regularizing by the total amount of all times by equation (13).

$$a_{ij} = p\left(S_j|S_i\right) = \frac{expected\ number\ of\ transitions\ from\ state\ i\ to\ j}{expected\ number\ of\ transitions\ from\ state\ i} \quad (14)$$

To calculate the assessed possibility for a given transition from i to j equation (14) is used. It was taken at a particular point t in the detected sequence. The possibility for each specific point t was identified, the total amount for transition I to j can be calculated.

IV. EXPERIMENTAL RESULT AND ANALYSIS

A. Dataset description

In order to estimate the performance of human action recognition using two datasets: HMDB-51[7] and UCF-50[18]. The UCF-50 dataset is one of major action datasets that contains 6681 videos including 50 human action categories. The videos are gathered from the YouTube. The videos which are used is divided into 25 groups covered by 4 action clips from each group of human action. The resolution of all the videos is 320 × 240.

The HMDB-51 action dataset consists of 6849 videos, for each action it contains 70% training and 30% testing. The videos are collected from various resources such as movies, YouTube and other publicdatabases. There are different types of actions and some videos are with lower quality. The resolution of all the videos is 320x240. An example of the HMdb-51 dataset is shown in Figure: 3.

B. Results and Discussion

The action from the dataset HMDB-51 and UCF-50 were tested and the results are evaluated. Due to the poor quality of videos on HMDB-51 the accuracy for each action class on HMDB-51 dataset was less while comparing with UCF-51 dataset. UCF-51 dataset produces good results for each action class. Table. 1 and Fig.4 shows the accuracy results of each action class from the two datasets. In HMDB-51 dataset consists of 6849 videos and UCF-50 dataset consists of 6681 videos for each action it contains 70% training and 30% testing.

In proposed system, the performance is evaluated by testing new images. TP, TN, FP, FN values are evaluated by analyzing the actual and predicted class label which in terms contribute to the evaluation of accuracy. The accuracy is the proportion of the total number of predictions that were correct. Accuracy is calculated using equation (15).

$$Accuracy = \frac{TP+TN}{P+N} \quad (15)$$

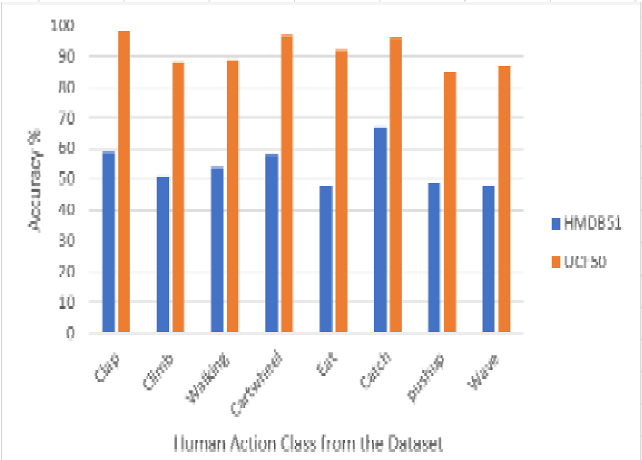

Figure 3: Comparison of Accuracy between HMDB-51 and UCF-50 dataset.

Table 1:Performance Analysis on Different Datasets

Action Class	HMDB-51 Accuracy (%)	UCF-50 Accuracy (%)
Clap	59.68	98.65
Climb	51.77	88.76
Walking	54.01	89.15
Cartwheel	58.98	97.20
Eat	48.54	92.40
Catch	57.28	96
Pushup	49.05	85.65
Wave	48.92	87.27

In the proposed work, human action is recognized using CNN and HMM and overall performance is evaluated and compared. CNN is a deep learned feature extraction in which each feature is extracted from each layer. In HMM hand crafted features such as canny edge detection and optical flow is used to extract the features. Table. 2 and Figure 4: shows the result for the overall performance of human action recognition.

Table 2. Comparison Between Two Approaches

Action Class	CNN Accuracy (%)	HMM Accuracy (%)
Clap	98.02	92.12
Climb	88.24	84.79
Walking	89.93	87.40
Cartwheel	97.20	94.45
Eat	92.05	89.25
Catch	96.05	91.15
Pushup	85.45	82.56
Wave	87.28	84.94

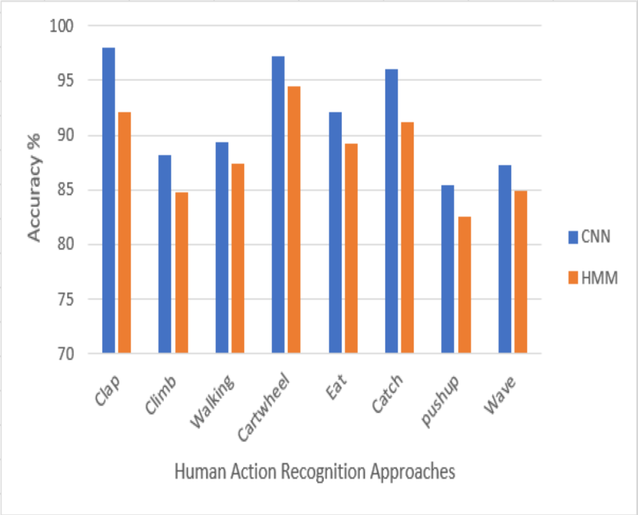

Figure 4: Comparison of Accuracy between CNN and HMM

Figure: 5 shows the comparison of human action recognition using Hidden Markov model. PalwashaAfsar et al. [2] has been stored all the feature vectors in a codebook where each symbol corresponds to an action by using Vector Quantization (VQ). For the training phase, the model parameters of HMM are tuned well for description of an action to achieve high results and obtained 84.96% accuracy. M. Ahamad et al. [16] has been mapped the estimated angle features from the time sequential activity video frames into code words to generate a sequence of discrete symbols for a hidden markov model of each activity and obtained 87.5% accuracy. Figueoa-Angulo et al. has been labeled the activities by means of a Compound Hidden Markov Model (CHMM). Each separate Linear Hidden Markov Model (LHMM) has motion information of a human activity and obtained 59.37% accuracy. In my proposed work, HMM is used for training and maximum likelihood estimation for testing and obtained 89.68% accuracy.

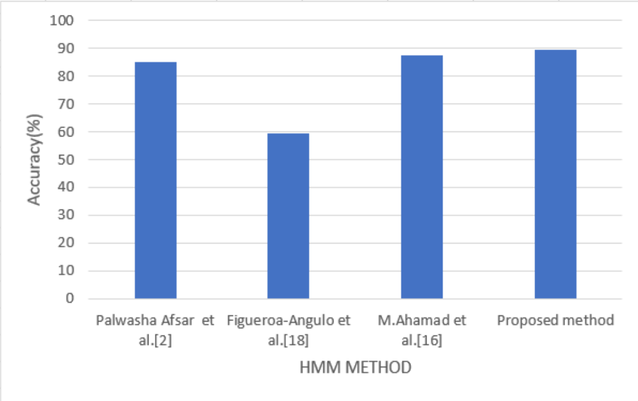

Fig. 5 Comparison of Accuracy between Different Method used in HMM

V. CONCLUSION

Comparative analysis is done for activity recognition using CNN and HMM. In the proposed work, multi-frame averaging method used for extracting RoI and forwarded to both CNN and HMM for comparative analysis. The features are extracted using CNN and hand-crafted features such as shape and optical flow based feature extraction. The feature extracted from CNN are used for Multiclass SVM classification. The hand-crafted features are extracted and fused together using Gradient fusion and used HMM for classification. The proposed approach was tested on HMDB-51 and UCF-50 dataset and the results were analyzed. In the future work, human action can be recognized for multiple person scenario or event detection and also other classifier will be used for evaluation.

REFERENCES

[1] M. Ahmad and Seong-Whan Lee. “HMM-based Human Action Recognition Using Multiview Image Sequences”. *IEEE 18th International Conference on Pattern Recognition*, vol. 4, pp. 874-879, 2006.

[2] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. “Large-scale Video Classification with Convolutional Neural Networks”. *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2014.

[3] . Corinna Cortes and Vladimir Vapnik. “Support-Vector Networks”. *Machine Learning*, vol. 20, pp. 273–297, 1995.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. “Behavior recognition via sparse spatio-temporal features”. *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, 2006.

[5] Figueroa-Angulo J., Savage J., Bribiesca E., Escalante B. and Scucar L. “Compound Hidden Markov Model for Activity Labelling”. *IEEE International Journal of Intelligence Science*, vol. 5, pp. 177-195, 2015.

[6] Fu Jie Huang and Yann LeCun. “Large-scale Learning with SVM and Convolutional Nets for Generic Object Categorization”. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006.

[7] Imran N. Junejo, Khurrum Nazir Junejo and Zaher Al Aghbari.

"Silhouette-based human
[8] action recognition using SAX-Shapes".*Springer*, vol. 30, pp. 259-269, 2014.
[9] Jie Yang, Jian Cheng and Hanqing Lu. "Human Activity Recognition based on the Blob Features". *IEEE International Conference on Multimedia and Expo,* pp. 358-361, 2009.
[10] Limin Wang, Yu Qiao, and Xiaoou Tang. "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors". *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 7–12, 2015.
[11] Maheshkumar H. Kolekar and Deba Prasad Dash. "Hidden Markov Model based human activity recognition using shape and optical flow based features". *IEEE Region 10 Conference (TENCON),* pp. 393-396, 2016.
[12] NavidNourani-Vatani, Paulo V. K. Borges and Jonathan M. Roberts. "A Study of Feature Extraction Algorithms for Optical Flow Tracking . Australasian Conference on Robotics and Automation". *Australian Robotics and Automation Association*, 2012.
[13] PalwashaAfsar and Paulo Cortez. "Automatic Human Action Recognition from Video Using Hidden Markov Model". *IEEE 18th International Conference on Computational Science and Engineering,* pp. 105-109, 2015.
[14] Sheng Yu, Yun Cheng, Songzhi Su, Guorong Cai, and Shaozi Li. "Stratified pooling based deep convolutional neural networks for human action recognition". *Multimedia Tools and Applications*, vol. 76, pp. 13367–13382, 2016.
[15] Xiaojiang Peng, LiminWang, XingxingWang, and Yu Qiao. "Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice", *Elsevier*, vol. 150, 2016.
[16]] Xin Yuan and Xubo Yang. "A Robust Human Action Recognition System Using Single Camera". *IEEE International Conference on Computational Intelligence and Software Engineering*, pp. 1-4, 2009.
[17] Md. Zia Uddin, Nguyen Duc Than and Tae-Seong Kim. Human. "ActivityRecognition via 3-D joint angle features and Hidden Markov models". *IEEE International Conference on Image Processing Electronics and Telecommunications Research Institute(ETRI) Journal*, pp. 713-716, 2010.
[18] Zhenzhong Lan, Shoou-I Yu, Ming Lin, Bhiksha Raj, and Alexander G. Hauptmann. "Local Handcrafted Features Are Convolutional Neural Networks". *International Conference on Learning Representations*, pp 43–56, 2016.
[19] http://crcv.ucf.edu/data/UCF50.php