International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869, Volume-3, Issue-4, April 2015

Human Action Recognition using Support Vector Machine and K-Nearest Neighbor

Sonali, Ashok Kumar Bathla

***Abstract*— Recognizing human actions from video sequences is termed as Human Action Recognition. It has many important applications like video surveillance, patient monitoring, human computer interaction, dance choreography analysis, analysis of sports events and entertainment environments. In this paper the work of the enhancement of human action recognition of the videos with the help of a Hybrid technique has been demonstrated. It proceeds in a step-wise approach. Firstly the train feature database is created to store the features. The testing of videos is then done on basis of support vector machine and K-nearest neighbor classifiers. The K-NN classifier here works by taking up Euclidean distances between the test and train features. Also Region of Interest is extracted by highlighting the boundary box to recognize the action and specific action labels are applied in the videos and the Non-ROI is enhanced using the median filter. The results obtained are quite significant and are analyzed on the public benchmark Weizmann dataset, which contains examples of bending, running, walking, skipping, and hand-waving of two types with both one and two hands, with nine actors performing these actions.**

***Index Terms*— action recognition, classification, support vector machine, nearest neighbor, bag of visual words**

I. INTRODUCTION

In today's life style Computer is a very essential and important machine. Manually work done by people is now done by Computer in very less time, efficient and with more accuracy. With the fast pace of development in the field of human computer interaction, human and its activities also need a broader study in order to develop more human-computer friendly systems. Human actions are not merely due to the movement or motion of body-parts of a human-being, rather it is the depiction of one's intentions, behavior and thoughts. "Action Recognition" as the term itself is self-suggesting, it is the recognition of an activity or action by using a system that analyzes the video data to learn about the actions performed and uses that acquired knowledge to further identify the similar actions[1]. Recognizing human actions and activities is a key-component in various computer applications like video-surveillance, healthcare systems, recognition of gestures, analysis of sports events and entertainment events. Human actions are not merely due to the movement or motion of body-parts of a human-being, rather it is the depiction of one's intentions, behavior and thoughts. "Action Recognition" as the term itself is self-suggesting, it is the recognition of an activity or action by using a system that analyzes the video data to learn about the actions performed and uses that acquired knowledge to further identify the similar actions. Human action recognition has been a popular research topic in recent two decades. Most previous works in this topic employed a frame-by-frame comparison to trained action models for classifying a newly arrived video sequence, which is computationally expensive due to the following facts:

- The consecutive frames in a video are correlated/similar in temporal domain; hence it is redundant to compare every frame for classification.
- In some cases, only a few frames in a video are sufficient for discrimination of basic actions [17].

Human activity recognition can serve many application areas, ranging from visual surveillance to Human Computer Interaction (HCI) systems. Visual Surveillance uses it as the video technology is becoming progressive, the visual surveillance systems have undertook a rapid development process, and have more or less become a part of our daily routine [11]. Human activity understanding can help to find fraudulent events such as burglaries, snatching, thefts , violent actions, etc. and can serve to track patients who need special attention (like identifying the well-being of a lonely person, detecting a falling person). Since, ubiquitous computing has increased the presence of HCI systems almost everywhere. A recently evolving thread is in the area of electronic games where we imitate the actions of a real world human being to create his avator on system. Due to substantial decrease in the cost of video capturing devices, videos have become a considerable part of the today's personal visual data. Automatic recognition of those data files, together with movies and other video clips helps information retrieval. Content based browsing and video recycling is aided with human activity recognition as in where the viewer is interested in only some specific parts of video archives, e.g. fast-forward to the next goal scoring scene. Human scientists analyze the impact of the entertainment industry especially the movies on the youth and the adolescents, human scientists use human actions recognition, in order to keep a record of both the positive and adverse influence on the young people, e.g. influence of smoking in movies on adolescent smoking. Gesture recognition, which is a sub-domain of action recognition, also operates over the upper body parts and recognition is applied. Thus it serves a lot for automatic understanding of sign language [6].

II. TYPES OF HUMAN ACTIONS

The study of human actions derives that the actions ranges from simpler actions to quite complex actions. With this

Manuscript received April 22, 2015.

Sonali, Research Scholar, CE Deptt., YCOE, Punjabi University, Patiala, India

Ashok Kumar Bathla, Assistant Professor, CE Deptt.,YCOE, Punjabi University, Patiala, India

intention we uncover the categories of human actions and activity types as :

- Single actor/single action
- Multiple actors/multiple actions

A. *Single actor/single action*

The first scenario covers up only for a single actor in the video. There are basically three key elements that define a single action:

- body postures
- relative ordering of the postures (2D/3D)
- speed of the body and adjoining body parts

We can formulate single action recognition as an embodiment of above mentioned three elements. The relative importance of the mentioned elements is based on the nature of the activities that we desire to find out (recognize). For example, if we want to differentiate an instance of a "bend" action from a "jump" action, the pose of the human body gives sufficient information. However, if we want to do the same between "run" and "jog" actions, the pose alone may not be enough, due to the similarity in the nature of these actions in the posture domain. In such cases, the speed information needs to be incorporated with the pose estimation measures. Various attempts have been made in action recognition literature try to model some or all of these aspects [4]. For instance, methods based on Spatio-temporal templates [9,20] pay attention to mostly the poses of human body, whereas methods based on dynamism focus their attention to model up the ordering of the poses in a broad context including more details [3]. As there is a shortage of benchmark datasets for working on human action recognition tasks, we merely focus upon single actor and single action movements. Figure-1 demonstrates different actions of drinking, walking, dancing and bowling *etc.*, one action taken at one time.

Figure-2.1: Examples of different actions of drinking, walking, dancing and bowling

B. *Multiple Actors/Multiple Actions*

In the second scenario, the case of complex activity recognition is considered, where the action units are composed over time and space and the viewpoints of the subjects are varying frequently. Because composition makes so many different actions possible [13], it is not reasonable to expect to possess an example of each activity [18].

III. RELATED WORK

A considerable amount of previous work has addressed the question of human action recognition. Aggarwal et al. (2011) [1] presented an overview of the current approaches used for human activity recognition and also yielded from it that, there is a diversity results and they vary with varying conditions of approach and dataset. A summarization of previously existing methodologies with both their pros and cons is done. The various application areas of human activity recognition like surveillance, monitoring services at public places, patient monitoring have been discussed with viewing the activities present as either of a group or individual. It has presented the significant program in the area of human activity recognition after a intensive study of all the existing issues and challenges posing to the process. Blank et al. demonstrates an approach to represent actions as space-time shapes and shows that such a representation contains rich and descriptive information about the action performed. The quality of the extracted features is demonstrated by the success of the relatively simple classification scheme used (nearest neighbor classification and Euclidian distance. It clearly specifies the various advantages as provided by using the proposed approach in terms of partial occlusions, non-rigid deformations, significant changes in scale and viewpoint, high irregularities in the performance of an action and low quality video[3]. Brendel et al. presented that certain human actions could be efficiently represented by short time-series of activity codewords [4]. In addition, those codewords may represent objects that people interact with while performing the activity. It has been observed in this with small computation times, we outperform the on the benchmark datasets. Deng et al. proposed a low complexity scheme for Region-of-Interest extraction [5]. It has been done on basis of the perceptual characteristics and the information theory. The spatial and temporal computation is done by identifying inter-frame correlation and pixel correlation. Kaghyan et al. presented an approach for classifying human activities by using mobile devices. The method was based on K-nearest neighbor algorithm. A single accelerometer is used to recognize activities based on the cell phone movements. The entire algorithm for the purpose of classification and its processing was elaborated in this paper in details [7]. Also it explains how sensors are used for the process and different accelerations are undertook to demonstrate the recognition task on the smartphones. Ke et al. extensively surveys the current progresses made toward video-based human activity recognition. Three aspects for human activity recognition are addressed including core technology, human activity recognition systems, and applications from low-level to high-level representation. In the core technology, three critical processing stages have been thoroughly discussed mainly the human object segmentation, feature extraction and representation, activity detection and classification algorithms. In the human activity recognition systems, three main types are mentioned, including single person activity recognition, multiple people interaction and crowd behavior, and abnormal activity recognition. The domains of applications are discussed in detail, specifically, on surveillance environments, entertainment environments and healthcare systems and the challenges associated with it [8]. Schuldt et al. demonstrated how local space and time features can be used for identifying complex motion patterns and activity in videos. The representation of motion pattern using this approach is robust to varying scales, velocities and frequency. It also explained how action like running and jogging could be distinguishly recognized from videos using

SVM [14]. Tran et al. proposed an efficient method for learning of motion difference features from actions or activities in videos. This took a base from the fact that a single visual word was difficult to assign to each subset of a frame. It used Gaussian Restricted Boltzmann Machines for recognition purpose. Firstly the difference between two frames of a video is obtained by background subtraction, which removes the irrevelant shapes and background parts for actions learning and recognition task. It reported that the usage of Gaussian restricted Boltzmann machine [15] gave a good performance in benchmark datasets like Weizmann (98.8%) and KTH (88.8%). Wang et al. proposed the usage of high-level action units to represent human actions in videos and based on such units, a novel sparse model is developed for human action recognition. Three interconnected steps [17] were carried out. Firstly, a context-aware descriptor named locally weighted word context is used. Secondly, from the statistical values of the context-aware descriptors, learning of the action units using the matrix factorization was done, which leads to a part-based representation and encodes the geometrical information of the video sequence. These units effectively bridge the semantic gap in action recognition. Lastly in order to suppress the noise, a sparse model was used with joint l2, 1-norm to normalize the data. Results have shown up that it outperforms with various datasets like KTH, UT-interaction, UCF, Youtube etc. [15]. Zou et al. presented a method that learns features from spatio-temporal data using independent subspace analysis. The algorithm to large receptive fields by convolution and stacking and learn hierarchical representations. Experimental results are obtained from KTH, Hollywood2, UCF sports action and YouTube datasets using a very standard processing pipeline. The results have been obtained using the same parameters across four datasets [20], which are consistently better than a wide variety of combinations of methods.

IV. GENERAL FRAMEWORK

Generally speaking, the task of human activity recognition can be divided into three levels as shown in figure 4.2 comprising of pre-processing and object segmentation, feature extraction and representation and activity detection and classification as explained in the next sub-sections.

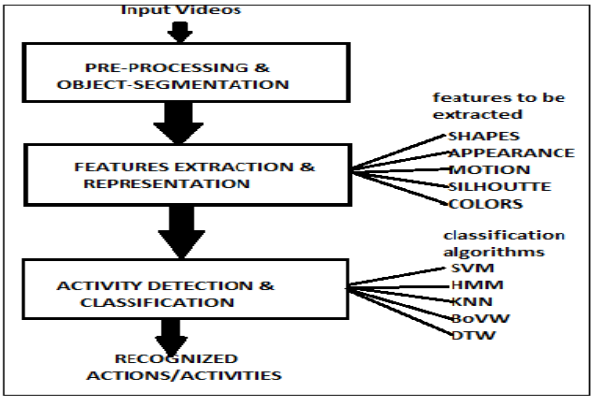

Figure 4.1:–Human Activity Recognition Process

A. *Pre-processing & segmentation*

The pre-processing stage involves the extraction of frames from the video as most of the previously done work in the field of human activity employs a frame-by-frame processing. Segmentation is done to extract the target object from the frames depending upon the camera mobility from which the videos were captured. The Object Segmentation has been shown in figure 4.2. For the static cameras, the camera alignment is fixed in a specific position and angle. As the background never moves, one can use the background subtraction method, wherein the current image of the background image is subtracted to get the required foreground object. On the other hand, contrasting to the simplicity of static camera segmentation, moving camera segmentation is quite challenging due to the fact that both the motion of the target object and the camera orientation and background keeps varying. The most common method for segmenting such videos is identifying the temporal difference between the consecutive frames [8].

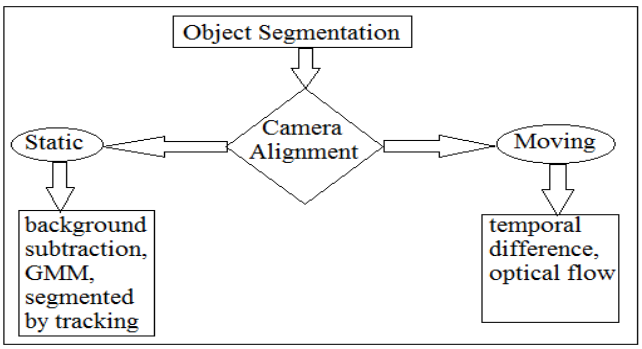

Figure 4.2- Object Segmentation techniques

B. *Feature Extraction and Representation*

Once the region of interest (ROI) is obtained from a frame, feature extraction is done where features like color, silhouette, shape are extracted. In a video sequence, the features that capture the space and time relationship are known as space-time volumes. The features could be space-time information, body modeling, local descriptors, gait pattern, silhouette etc. as shown in figure 4.3 [6].

Figure 4.3-Feature Extraction using descriptor [8]

C. *Activity Detection and Classification*

Then is the classification which helps to recognize the human activities on basis of the features extracted. The classifiers use to recognize and classify the actions are Support Vector Machine[14], K-Nearest Neighbor[7], Dynamic Time Warping, Hidden Markov Model etc. as shown in figure 4.4. the next sub-section explains how SVM and K-NN are used in this task of Human Action Recognition.

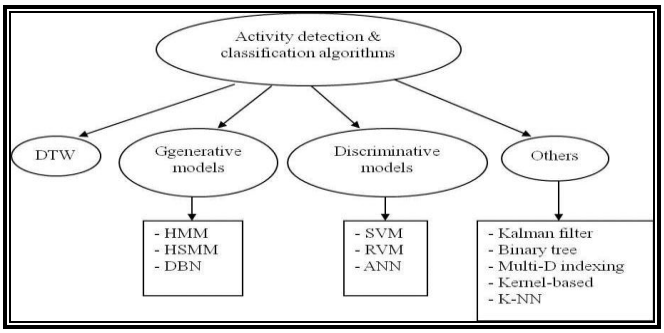

Figure 4.4-Categories for activity detection and classification algorithms [8]

a) Support Vector Machine

SVM has a higher generalization capability and provides high accuracy. SVM creates a hyperplane for classifying the data into a high dimensional space for separating the data with different labels. On each side of the hyperplane created initially, two separate hyperplanes are created. SVM tries to find that hyperplane which maximizes the distance between the two parallel hyperplanes. A wisely done separation means largest distance between the hyperplane and the nearest training data point of any class [14].

b) K-Nearest Neighbor

K-NN classifier measures the distance between the image or frame representation obtained from an observed sequence of video and the training set. It is amongst the simplest of all machine learning algorithms wherein the features extracted are classified by a majority vote of its neighbors, with the feature or object being assigned to that class, which is the most commonly occurring in its k nearest neighbors. The working of K-Nearest Neighbor starts by choosing a closest neighbor of the element that needs classification to be done. As explained in the figure 4.5, when k=1, we assign the element in green to the banana class which is in yellow color on the basis of Euclidian distance measures as shown in figure 5.5. Likewise, when we take k=3, the element to be classified is assigned to apples class on basis of majority voting carried out as now the neighbors of the element to be classified are in yellow and red color [7].

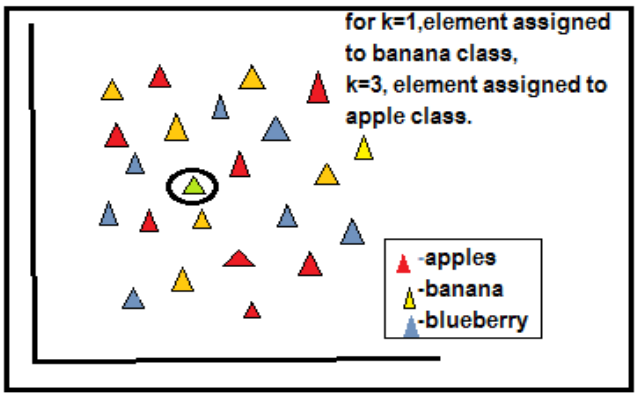

Figure 4.5 K-NN classification

V. ALGORITHM FOR IMPLEMENTATION

The proposed work of this system is given below in the following steps:-

Step 1: Create the dataset of the video that is to be tested for action recognition.
Step 2: Create the train_Feature_DB_file_32.mat to store the all features of human action on ROI part with the help of following:
Step 3: Read the sample .avi video that is be the tested from the directory .
Step 4: Click on radio button full or sample according to the choice.
Step 5: Click on testing sample video or full button according to the radio button.
Step 6: Apply SVM and KNN classifier to classify the features with the help of function svmclass & kNN_classifiers.
Step7: Extract the feature of ROI part of the video with where video is processed according to the frame and the size_ROI_x and size_ROI_y is marked.
Step 8: Enhance the Non-ROI part of the video with median filter to improve the human visibility of the video.
Step9: Insert the label on human actions.
Step 10: Calculate the Accuracy of the recognized human action and confusion matrix according to the full testing videos.

VI. RESULTS AND DISCUSSION

As mentioned earlier, this work is based upon a well defined graphical user interface to make the program easier to use. The following figure 6.1 is the graphical user interface that is used to recognize the human actions. The figure 6.2 is used to show the browsing of the input videos that is used to process for action recognition.

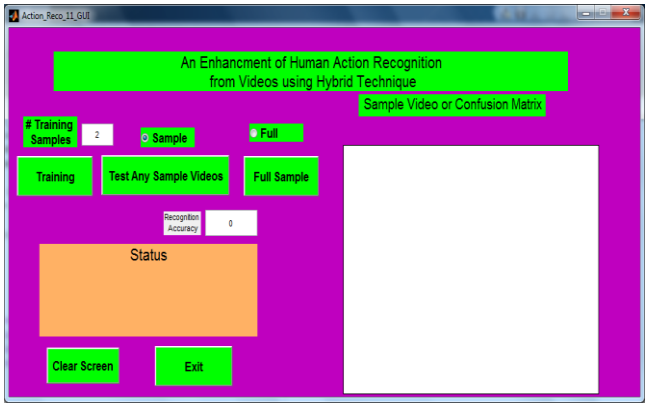

Figure 6.1 Starting Graphical User Interface

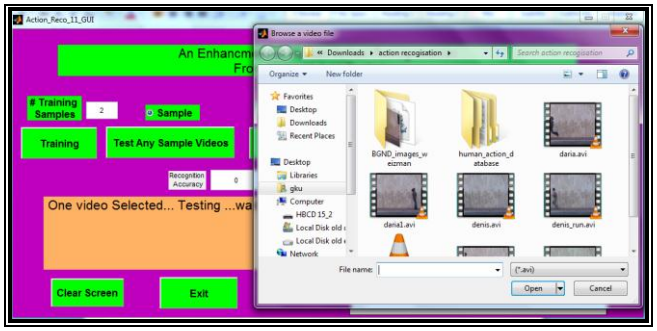

Figure 6.2 Browse the input video in Graphical User Interface

A. *Test videos*

It is a collection of different clips which are kept in a separate folder consisting of videos of distinct actions performed by different persons.

a) Single Sample Testing

In this, we have a button on GUI to browse single sample of a video to be tested at a particular instance of time. The figure 6.3 shows the sample output of test video recognizing the running action with 100% accuracy**.**

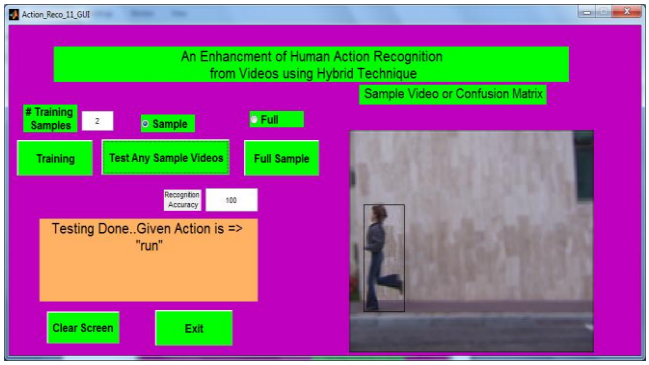

Figure 6.3 Sample output for single sample test video

b) Full Testing

Similar to the action performed for a single sample testing, here we have done full testing of the entire video clips contained in the dataset. The figure 6.4 defines the initialization step of full testing. The figure 6.5, figure 6.6 and 6.7 shows up the actions of skip, jump, walk respectively with labels as full testing proceeds.

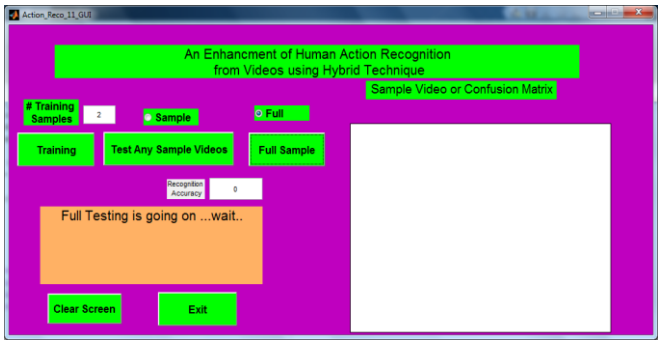

Figure 6.4 Initialization for the full test video

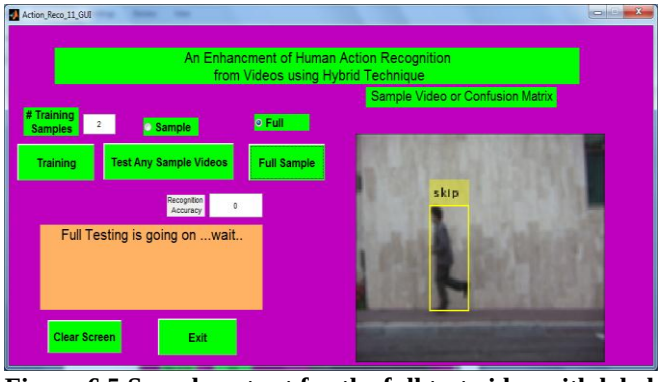

Figure 6.5 Sample output for the full test video with label "skip"

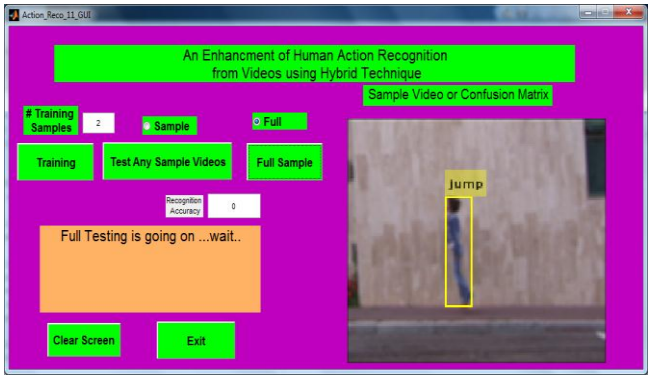

Figure 6.6 Sample output for the full test video with label "jump"

Figure 6.7 Sample output for the full test video with label "walk"

As the full testing is done, the accuracy and confusion matrix of the entire dataset is displayed in the confusion matrix as shown in figure 6.8 below.

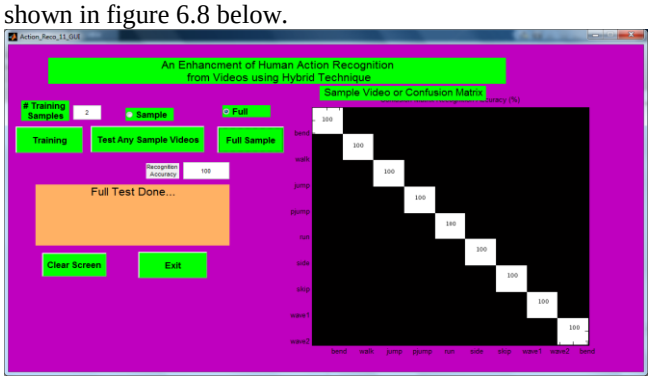

Figure 6.8 Sample output for the full test video with confusion matrix and accuracy

Table 6.1 demonstrates the results we have obtained from our approach applied. It should however be noted that it however it performed quite perfectly with a hybrid of SVM and KNN on only the actions of Weizmann dataset, it is expected the technique would perform well on other datasets also although with a low accuracy.

Method	Approach Used	Accuracy (%)
Our method	SVM + KNN	100.0
Tran et al.[15]	MD + Gaussian RBM+ NB	98.8
Zhang et al.[19]	pLSA + SVM	93
Niebles et al. [10]	pLSA + LDA	90.0

Table 6.1 Performance on Weizmann dataset, the results of [15], [19] and [10] are copied from the original papers.

VII. CONCLUSION

Although the performance of our method is comparable with other classical methods like Support Vector Machine (SVM) and K-Nearest Neighbor classifiers, the recognition rate is dependent on foreground and background extraction of the video. The foreground part includes the human action and background part include the static background of the video. In the action recognition systems it is difficult to enhance the background part of the actioned video and recognition of the action in ROI part of the video. There is sparse decoding data loss problem due to ROI and NOI-ROI region of the action detected video. We proposed a framework for human action detection in a video. There is video data set that we have to test and to find the Region of Interest and Non-ROI part of the video. The ROI part is extracted to detect the action of the human with Support Vector Machine and K-Nearest Neighbor classification and enhancement of Non –ROI part of the video with median filter. The accuracy of the recognized action is 100 % on each sample and the full testing of the videos for only the nine actions of Weizmann dataset. The results are promising but still due to owing lack of large datasets, much of work could be done to monitor patients using medical video clips. In addition, we can plan further to have a more discriminative approach for finding other actions than ones described in Weizmann dataset.

REFERENCES

[1] Aggarwal J. K., Ryoo M. S.,“ Human Activity Analysis: A Review”, Journal on ACM Computing Surveys (CSUR), 2011, Vol.43, No. 3, pp.1-47.

[2] Bengalur M.D., “Human Activity Recognition using Body Pose Features and Support Vector Machine”, Proceedings IEEE International Conference on Advances in Computing, Communications and Informatics, Mysore, 2013, pp.1970-1975.

[3] Blank M., Gorelick L., Shechtman E., Irani M., Basri R., “ Actions as Space-time Shapes” in Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV), Beijing, 2005, Vol. 2, No.1550-5499, pp.1395-1402.

[4] Brendel W., Todorovic S., “Activities as time series of human postures”, Proceedings of ECCV, Crete, 2010, Vol. 6312, pp.721-734.

[5] Deng J., Lu Z., Wen X., Wang L., Shao H., “Information Theory based Region of Interest Extraction Scheme with Perceptual Stimulus-Response Model”, IEEE 24th International Symposium on Personal, Indoor and Mobile Radio Communications: Services, Applications and Business Track, London, 2013, No. 2166-9570, pp. 3528-3532.

[6] Hong P., Turk M., Huang T., “Gesture modeling and recognition using finite state machines”, Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, 2000, pp.410–415.

[7] Kaghyan S., Sarukhanyan H., “Activity Recognition using k-nearest neighbor algorithm on smartphone with tri-axial accelerometer”, International Journal on Information Models and Analyses”, 2012, Vol 1, pp. 146-156.

[8] Ke S. R. , Thuc H. L. U., Lee Y. J., Hwang J. N. , Yoo J. H. , Choi K. H., “A Review on Video-Based Human Activity Recognition”, 2013, Vol. 2, pp.88–131.

[9] Le Q.Y., Zou W.Y., Yeung S.Y., Ng A.Y., “Learning Hierarchical Invariant Spatio-Temporal Features For Action Recognition With Independent Subspace Analysis,” in IEEE (CVPR), 2011, Issue No.1063-6919, pp.3361-3368.

[10] Niebles J., Wang H., Fei-Fei L., “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words”, International Journal of Computer Vision, 2008, Vol. 79, No. 3, pp. 299–318.

[11] Paul M., Haque S. M. E., Chakraborty S., “Human detection in surveillance videos and its applications - a review”, Springer EURASIP Journal on Advances in Signal Processing, 2013, pp.1-16.

[12] Ramanathan M., Yau W.Y., Teoh E.k., “Human Action Recognition with video data: Research and Evaluation Challenges ”, IEEE Transaction on Human-Machine Systems, 2014, vol.44, no.5, pp.650-663.

[13] Ryoo M.S., Aggarwal J.K., “Semantic Representation and Recognition of Continued and Recursive Human Activities”, International Journal Of Computer Vision, 2009, Vol.82, no. 1, pp. 1-24.

[14] Schuldt C., Laptev I., Caputo B., “Recognizing Human Actions: A Local SVM Approach”, IEEE International Conference on Pattern Recognition, 2004, Vol 3, pp. 32-36.

[15] Tran S. N., Benetos E., Garcez A. D., “Learning Motion-Difference Features using Gaussian Restricted Boltzmann Machines for Efficient Human Action Recognition”, IEEE International Joint Conference on Neural Networks (IJCNN), Beijing, 2014, pp.2123-2129.

[16] Vemulapalli R., Arrate F., Chellappa R., “Human Action Recognition by Representing 3D Skeltons as points in a lie group ”, IEEE Conference on computer vision and

[17] Wang H., Yuan C., Hu W., Ling H., Yang W., Sun C., “Action Recognition Using Nonnegative Action Component Representation and Sparse Basis Selection”, IEEE transactions on image processing, 2014, Vol. 23, No. 2, pp.570-581.

[18] Yang M., Lv F., Xu W., Yu K., Gong Y., “Human Action Detection by Boosting Efficient Motion Features” IEEE 12th International Conference on Computer Vision Workshops (ICCV), Kyoto, 2009, pp. 522-529.

[19] Zhang J., Gong S., “Action categorization by structural probabilistic latentsemantic analysis” ELSEVIER, Computer Vision and Image Understanding, 2010, pp. 857-864.

[20] Zou W.Y., Le Q.Y., Yeung S.Y., Ng A.Y., “Learning hierarchical invariant Spatio-temporal features for action recognition with independent subspace analysis,” in IEEE (CVPR), 2011, No.1063-6919, pp.3361-3368.