International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1874-1879

Real Time Action Recognition in Surveillance Video Using Machine Learning

Abdulrahman S. Alturki[1*] and Anwar H. Ibrahim[2]

[1, 2] *Department of Electrical Engineering, College of Engineering, Qassim University, Qassim, Saudi Arabia.*

Abstract

Human gesture identification plays a crucial and key role in the surveillance and security domains. This technique is most wanted in today's world to identify the culprits or specific people over the surveillance cameras. In this proposed method the action identification is aided through the machine learning technology. Initially the frames of the subject under focus are segmented and modeled by the Gaussian modeling. The entire process of feature extraction and quantizing are done considering the characteristics of the area of interest. The proposed method has two phases testing and training phase using three datasets for validation and KNN classifier for classification. The output of the proposed algorithm is analyzed and is said to be have the correct gesture identification with accuracy rate of 95.568%.

Keywords: KNN Classifier, datasets, Gesture identification, Prewitt filter, GMM modeling, machine learning, feature extraction, multiple view points

I. INTRODUCTION

For the past two decades the domains such as Machine learning and computer vision has set an challenging goal to recognize the actions of human in an autonomous mode. Now the approach has its extensive applications in many emerging fields like deep learning, medical industry, Security systems, intelligence surveillance, etc. Thus it is in the limelight and draws more attention among the researchers at present[1]-[3]. The identification of human action through computer technology has its strong implementation in the real world applications like Storing Big files[4], Action identification[5], Indexing[6] and securing the videos[7], etc. The critical and important part of this technology is the interaction between the machine and the human. Visual signal plays an crucial rule in the recognition of the human actions and the communication between the computer and human. The recent developed techniques involves manual sampling of the signal before digitizing in the computer[9]. There are some practical impossibilities and difficulties in setting the starting and end portion of the sequence samples. Hence an algorithm was developed to automate the action division in an image sequence. The actions of the subjects in an image sequence can vary based on the pattern, pose, mobility, etc. These parameters are yet an challenging issues which affects the image properties like luminance, chrominance, Background sequence, etc. The most important key point is the view point dissimilarity as the HAR approach are based on the visual signals which is recognized from the single view of capturing. For the entire process starting from training and ending with the testing same camera is used to capture the views. But the real time applications cannot have same setup in their process. This causes vigorous fall in the accuracy of the different views. The failure of the single view techniques occurs due to the hiding of a portion due to some inevitable obstructions. To overcome these limitations and to obtain the absolute image multiple cameras are used to capture the image. This led to the discovery of the term Multi-view action identification.

In this multi view approach the images are captured either in the form of 2D or 3D. In three dimensional capturing the object under focus is segmented into multiple views[9] and the mobility depiction is framed for identification of the movements. The model construction in this approach involves the use of geometrical shapes as patterns. The 3D approach is generally used in the real time applications such as Histogram shaping, optical patterns in 3D, Storage of action history, Skeleton representation in 3D[10], action patterns in spatial domain, etc.

The 3D representation has the advantage of improved accuracy rate over the two dimensional capturing but with the drawback of cost expensive. Hence it is not much preferred in the real time applications. But yet the three dimensional capturing proved have fair quality of construction as the feature extraction of the approach depends upon the multi view capturing[11]. The errors in this approach arise due to the lack of proper details which are lost during segmentation of the 3D modeling. In general the best three dimensional model is constructed with the overlapping of the views[12][13]. Thus for the good quality 3D representation ample amount of views are required.

With the advancement in the image capturing techniques three dimensional cameras are readily available to capture 3D view of the object in focus. Out of many devices 3D Microsoft

*Corresponding Author

kinects and the ToF is most preferred for their ability to conquer the obstructions in capturing and reconstructing accurate images. The 3D sensors also has certain drawbacks and limitations. The 3D sensors are designed to capture the anterior view of the object in focus which is also distracted by the reflecting lights from posterior end. This limits the quality of the captured image sequence and hence the 2D capturing is preferred over 3D capturing.

There are several researches and analytical studies on the multi view capturing of 2D images. But the important process are alone discussed in this section. The first method is based on the feature extraction. Here the visual signals are represented by the various views from the respective descriptors. Then using relevant classifier the identification of the action is done[14]. In the second method the image segmentation is based on an classification scheme. Either an universal or multiple classifiers are used for the purpose of image segmentation. These classifiers are trained properly for perfect capturing and their end results are clustered for the absolute 2D image. The third model involves the incorporation of the variation of deep learning algorithms based on the machine learning model. Here the distinct features of the deep learning model is used to extract the image details from the raw data for action identification.

The proposed method doesn't require any distinct parameters or features for the processing of visual signals instead it exploits the spatial features from the multiple frames. The proposed method is capable of capturing both the universal and local boundary data and the specific distribution point data as well. In the proposed method Automation of the labeling of each action sequence at its start and end frame is done. Thus, at the reduced cost the sequential actions are identified effectively without any obstructions caused intentional or unintentionally. In this proposed model the identification of the multi view model is done with the help of the unalterable features of the visual signal.

II. STUDY BACKGROUND

The initial works in the action identification from the sequence of image is based on the flow of the visual signals [15][16]. Then the gesture recognition based on the secular-spatial domain came into existence [17]. The manipulation done to control the flow of visual signals is utilized to build the action frames which are used to track the gesture identification. The features present at the partition of each action segment are very subtle to the noise. This caused the extraction of high-resolution features from the real time signals with more noise. Hence to obtain the highest accuracy rate datasets are used [18]. Generally, Weizman data sets are used for the validation but for the signals with noise KTH datasets are used. This dataset is efficient in handling the low-resolution noises and geometrical attacks [19]. The two-dimension vectors are extended in the spatial domain to act better as 3D featured vectors. The action identification of 3D descriptors is more effective than the 2D descriptors because of the boundary data of the pixels [20].

By congregating the perceived pixels of the interest an glossary containing the prototypes can be made. To discover and unearth the fascinating pixels at the denser region a method was proposed by Dollar et al (2005)[21].The BOW approach was suggested for this process but it rarely samples the data which is an major drawback of the process. To enhance the feature of the action identification an composite model with the collection of the constellated features was proposed[22]. The enhanced features were strong against the obstruction caused to capture the image. These parameters focused on the local pixel location rather than the global pixel recognition. The works carried out so far did not had any consideration for the noise in the visual signals. Hence new methods were discovered to handle the obstructions and noises in the identification of gestures from the captured image.

The earlier studies reveal that all the gesture identification had a common consideration of the static camera without any motion capturing. But the gestures and poses vary with the viewpoints and the orientation of the camera angle. Some method involves training multiple classifiers for each view point or vice versa as single universal classifier for all viewpoints [22][23]. These considerations are used to outstretch the single view point towards the multiple views of the image capturing. All the extracted features and the classifiers trained contributes to the performance efficiency of the technique.

Later to study about the gesture of the images motion capturing was discovered by Lu et al (2012)[24]. This approach has the demerit of disfigured background when there is more than one subject in the focus. To overcome this issue and to ensure the accuracy many methods were proposed over the motion capturing and multi view dataset to identify the gestures in 2D and 3D. The above discussed techniques have performance degradation with the modified classifiers or the viewpoints. but the proposed technique is durable against the modified multiple viewpoints from different angle of orientations and produce an effective gesture identification with at most accuracy.

The proposed model is an 3D approach of action identification. In this approach the features are extracted from the available ample multi-view frames and the features are clustered to from the resulting action. For this purpose, the feature classifiers are used in the online training phase. Based on the requirement either single or multiple classifiers can be used to divide the features.

III. PROPOSED MODEL- AN OVERVIEW

The proposed method is a new action identification technique which is more durable against the change of frame, dimension, Illumination, etc. The model is processed at two phases namely training and testing. The training phase is an offline phase which deals with the extraction of features regarding fascinating information. Here the vector features required for the extraction are defined properly and are scaled down to

decrease their aspects. Then these vectors are accumulated and stocked in the database.

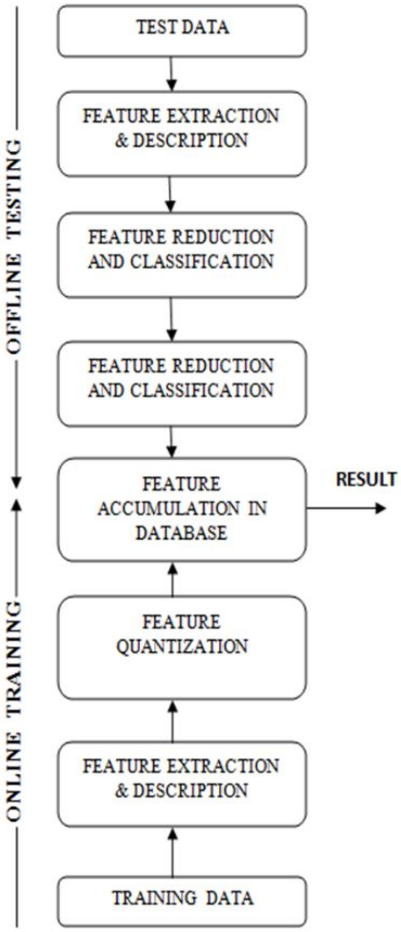

Fig.1. Proposed model flowchart

The training phase of the proposed method is considered as an online method. This phase also involves the extraction and labeling of the feature vectors. The scaling of these vectors is based on the derived histogram model from their features. Then the resultant vector is compared with the database to identify the actions. The detailed flowchart of this method is given in the Fig.1.

The Assessment of the obtained results id done with the different datasets which are explained in the following sections.

IV. ANALYSIS OF DATASETS

The datasets are used to assess the features extracted from the captured images. In 2004 KTH datasets were introduced by the Royal institute of technology. It was considered as the largest dataset to assess the video signals with different backgrounds. This dataset has six different action phases of human gesture. The analysis of each phase was done with the 25 experiment action sequences with 4 different backgrounds. Thus, it has total of 600 videos in the dataset performed by single subject. later the Weizmann institute created a new dataset with its name as Weizmann dataset. This dataset has videos of 9 subjects with 10 phases of actions. The third dataset used in this proposed model is Multiple camera human gesture video which was discovered in 2010. The videos in this dataset are captured by multiple cameras in multiple views of 14 subjects with 17 phases of action captured by 8 cameras.

The assessment of the result is done by comparing the features of the extracted output with the available samples from three different dataset to identify the human gesture.

V. PROCEDURE FOR EXTRACTION OF FEATURES

In this section the procedure for the extraction of feature from the quantized results are discussed in brief as given in Fig.2.

In the perception and identification of the gesture from the images by excluding the noise from the images is a hard and tough process. It involves multiple steps of works to be done. The first step includes the building of background scenarios using the GMM (Gaussian mixture model) [25]. Then to separate the subject from the scenario.

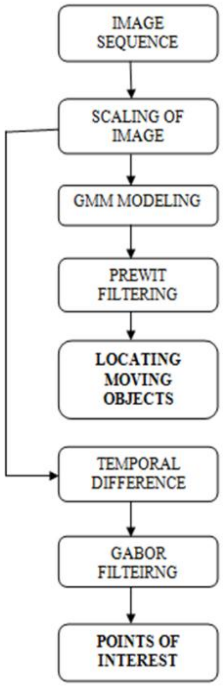

Fig.2. Procedure for extraction of features

Prewitt filtering is used. The hardest part in this phase is the construction of the background scenario in detecting the complicated human gesture. The steps to carry over the GMM modeling to build the background is explained as flowchart in Fig.3. The pixel intensity is not uniformly distributed throughout the image. Variations are there depending upon the color and illumination properties. Hence to build the background scenario GMM modeling is used as it has the uniform distribution through the entire function. Due to the presence of different objects and scenarios at the background uniform interval between each pixel variation exist.

To reduce the noise and to eliminate the ripples the frames are passed through the low pass filtering. Then the uniform pixel variation is applied by creating the boundary difference between each pixel. Then the modeling algorithm is applied to find the edges in the pictures using corresponding edge detectors. Thus, the entire background is updated and results in a GMM modeled scenario. The prewitt filtering forms a bounded box to identify the gestures. These bounded boxes are placed according to the gesture sequence. For instance, the actions done by hands are highlighted by the bounded box placed around the hands. Depending upon the placement of the bounded boxes by prewitt filters the area of interest are focused and the Gabor filter is used to separate the area of interest from the image. Also, this filter identifies the variation in intensities in the image. This is done by using the dollar detector algorithm. But there are certain limitations in this technique like the elimination of the movements, False identification due to the interference of noise present in the image, misleading background detection and weak identification of the slow movements.

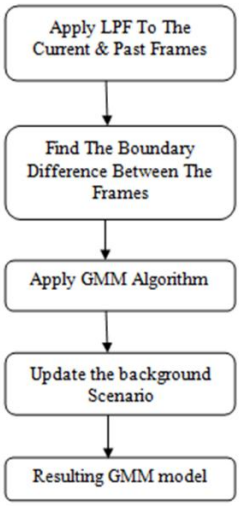

Fig.3. GMM modeling flowchart

Gabor filtering is a linear method to identify the edges in an image. This method works based on the frequency domain which is similar to the visual signals. This method is highly referred for the depict and to bias the texture of an image. When used in spatial domain it is modulated by the Gaussian function using sine wave. By combining the bounded boxes produced by prewitt filter, Gabor filter isolated the area of interest. This is used to depict the local and characteristic properties of the human gesture.

VI. FEATURE QUANTISING AND CLASSIFIERS

At the end of the training phase in offline mode the features quantized are stored in the form of vectors. In the online mode of testing phase these vectors are tested against the testing data and classified based on the reference classifiers. For the classification purpose classifiers like Gaussian classifier, Neighbor classifier and the mean classifiers are used.

Neighbor classifier is used to identify the gestures by the manipulation of the distance between the vectors and testing data. Here for the discrimination purpose voting is considered. The object which won major votes is classified against the neighbor classifier. The only limitation in using this classifier is that the computation time is higher when compared to that of the remaining classifiers. This classifier is considered for the proposed model.

VII. ANALYSIS OF RESULTS

The algorithm was developed in MATLAB and tested against the series of gestures. The inputs were given in the form of video feeds which are in uncompressed form. The video frames are adjusted so as to suit all the three datasets taken for assessment. The number of feature vector for training was set to 5. The scale limit was kept 6 so that the total s taken for consideration will be around 40. Hence for a single gesture around 1500 features are being manipulated. The input frame taken for the analysis is given in fig.4.

Fig.4. Input frame

The segmentation of the feed frame is done with the window resolution of 8x8, with the movement consideration of 2 pixels in each window. Total of 3 frames was used in each iteration

VIII. ASSESSMENT OF SUBJECT STABILITY

The stability of the subject is assessed through a cross validation technique which is used to calculate the rate of gesture identification. A group of frames consisting of subject of observation is selected as the test data by the dataset, and the remaining frames are considered for training phase. The iteration of each selected frames in the dataset is done to ensure the occurrence of all frames. The obtained recognition rate in all the iteration was 95.568%. Thus, all the gestures for the selected test subject was accurately identified by the proposed algorithm.

IX. ASSESSMENT OF THE VISUAL STABILITY

For this assessment the same set of procedures is followed as given in the above section. 5-6 gestures from a single point

view is selected to test against the dataset database. All the identification rates are calculated and recorded. The average of the identification rate is taken for the assessment result. The identification rate from single view is 81.56%. The same steps are followed for the multi view frames and the rate of identification is obtained as 90.897%. From this assessment result it is concluded that the identification through multi view point is more stable and accurate when compared to the single view. For the purpose of simplified algorithm single view approach is used in the proposed algorithm. The gesture of the feed frame was identified as running by the algorithm and displayed on the console.

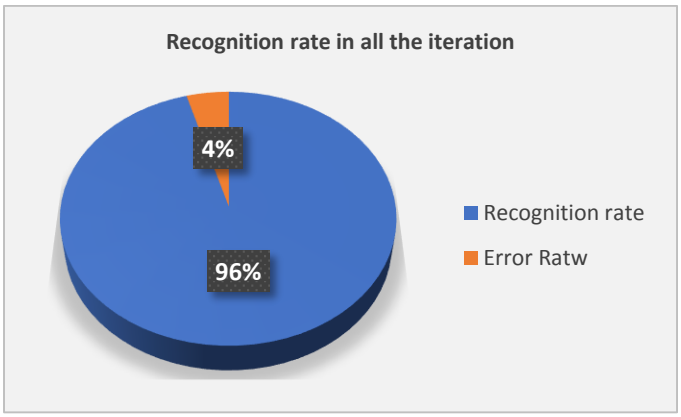

Figure 5. Recognition rate in all the iteration

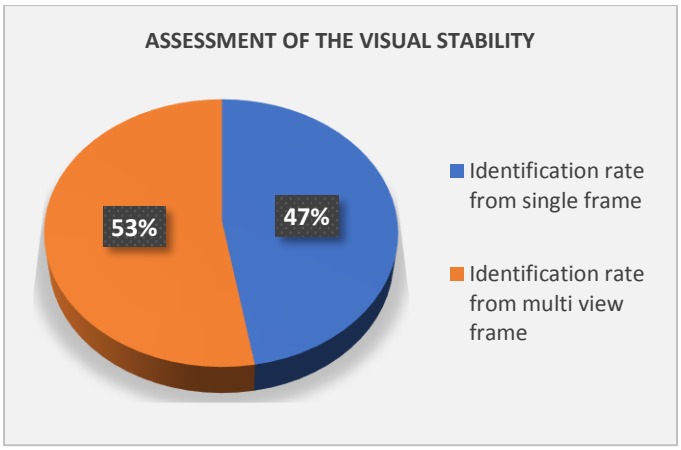

Figure 5. Recognition rate in all the iteration

X. CONCLUSION

The proposed method of gesture identification involves the real world applications and implementations. This method has an added advantage of finding the boundaries of each frame and segmenting them automatically. The system consists many visual features to identify the gesture from multiple view points. The features are extracted from the boundary conditions, universal classifiers and characteristics of the frame which are modeled using GMM algorithm. The modeling involves the identification and isolation of the area of interest by prewitt and gabor filters. By cross testing the extracted features and validating against the dataset the gesture of the subject under study is identified at the rate of 95.568% accuracy.

The future work of the proposed algorithm involves the implementation with the 3D images provided with ample amount of viewpoints from multiple cameras.

REFERENCES

[1] Alexandros.A., Jose.R.Lopez, Pau.C.Perez, Francisco Florez, "Evolutionary Joint Selection to Improve Human Action Recognition with RGB-D devices" Expert Systems with Applications, Elsevier Volume 41, Issue 3, 15 February 2014, Pages 786-794.

[2] Liu H., Ju Z., Ji X., Chan C.S., Khoury M. (2017) "Study of Human Action Recognition Based on Improved Spatio-Temporal Features". In: Human Motion Sensing and Recognition. Studies in Computational Intelligence, vol 675. Springer, Berlin, Heidelberg.

[3] Carvajal J., McCool C., Lovell B., Sanderson C. (2016) "Joint Recognition and Segmentation of Actions via Probabilistic Integration of Spatio-Temporal Fisher Vectors" In: Cao H., Li J., Wang R. (eds) Trends and Applications in Knowledge Discovery and Data Mining. PAKDD 2016. Lecture Notes in Computer Science, vol 9794. Springer, Cham.

[4] Chen Chen, Roozbeh Jafari, Nasser Kehtarnavaz "UTD-MHAD: A MULTIMODAL DATASET FOR HUMAN ACTION RECOGNITION UTILIZING A DEPTH CAMERA AND A WEARABLE INERTIAL SENSOR" IEEE International Conference on Image Processing (ICIP), December 2015.

[5] Wang, H., Oneata, D., Verbeek, J. et al. "A Robust and Efficient Video Representation for Action Recognition" Int J Comput Vis (2016) 119: 219.

[6] Karen Simonyan, Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", Advances in Neural Information Processing Systems 27 (NIPS 2014).

[7] Hossein Rahmani, Ajmal Mian and Mubarak Shah, "Learning a Deep Model for Human Action Recognition from Novel Viewpoints", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

[8] Liu J., Shahroudy A., Xu D., Wang G. (2016) "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition" Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9907. Springer, Cham.

[9] Raviteja Vemulapalli, Felipe Arrate, Rama Chellappa "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group", The IEEE Conference on Computer Vision and Pattern

Recognition (CVPR), 2014, pp. 588-595.

[10] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, Farid Boussaid "A New Representation of Skeleton Sequences for 3D Action Recognition", Pattern Recognition in computer vision foundation, IEEE explorer, 2017 - openaccess.thecvf.com.

[11] Hossein Rahmani, Ajmal Mian, "3D Action Recognition From Novel Viewpoints", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1506-1515.

[12] Xiaodong Yang, Ying Li Tian, "Effective 3D action recognition using EigenJoints", Journal of Visual Communication and Image Representation Volume 25, Issue 1, January 2014, Pages 2-11.

[13] Z.Gao, H.Zhang, G.P.Xu, Y.B.Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition", Elsevier, Neurocomputing Volume 151, Part 2, 5 March 2015, Pages 554-564.

[14] Mengyuan Liu, Hong Liu, Chen Chen "Enhanced skeleton visualization for view invariant human action recognition", Elsevier-Pattern Recognition, Volume 68, August 2017, Pages 346-362.

[15] A.A.Efros, A.C.Berg, G.Mori, and J.Malik, "Recognizing action at a distance," in null. IEEE, 2003, p. 726.

[16] A.Fathi and G.Mori, "Action recognition by learning mid-level motion features," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.

[17] Alexander Klaser, Marcin Marszalek, Cordelia Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients", BMVC 2008 - 19th British Machine Vision Conference, Sep 2008, Leeds, United Kingdom. British Machine Vision Association, pp.275:1-10, 2008.

[18] BMVC 2008 - 19th British Machine Vision Conference, Sep 2008, Leeds, United Kingdom. British Machine Vision Association, pp.275:1-10, 2008.

[19] C.Schuldt, I.Laptev, and B.Caputo, "Recognizing human actions: a local SVM approach," in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3. IEEE, 2004, pp. 32–36.

[20] D.G.Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[21] P.Dollar, V.Rabaud, G.Cottrell, and S.Belongie, "Behavior recog- nition via sparse spatio-temporal features," in Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. IEEE, 2005, pp. 65–72.

[22] A.Iosifidis, A.Tefas, and I.Pitas, "Neural representation and learning for multi-view human action recognition," in the 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, 2012, pp. 1–6.

[23] J.Gall, A.Yao, N.Razavi, L.Van Gool, and V.Lempitsky, "Hough forests for object detection, tracking, and action recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 11, pp. 2188–2202, 2011.

[24] Y. Lu, Y. Li, Y. Shen, F. Ding, X. Wang, J. Hu, and S. Ding, "A human action recognition method based on Tchebichef moment invariants and temporal templates," in Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012 4th International Conference on, vol. 2. IEEE, 2012, pp. 76–79.

[25] Kuang-Pen Chou, Mukesh Prasad, Di Wu, Nabin Sharma, Dong-Lin Li, Yu-Feng Lin, Michael Blumenstein, Wen-Chieh Lin, Chin-Teng Lin "Robust Feature-based Automated Multi-view Human Action Recognition System", DOI 10.1109/ACCESS.2018.2 809552, IEEE Access, 2018.