International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

RECOGNITION OF HUMAN ACTION INTERACTION USING MOTION HISTORY IMAGE

P. SANGEETHA

M.Tech Student, Dept of Information Technology, Anna University- CEG, Tamil Nadu, India.

Abstract *- Human Action Recognition (HAR) is more important in the field of computer vision because of growing demands in many applications, such as surveillance environments, entertainment and healthcare systems. Therefore, many research attempts have been undergoing to accurately detect the human activities using data mining technique. The action recognition process involves extraction of data from a video and to detect the final action using classifier. The recognition rate is affected by viewpoint changes, lighting, partial occlusion and background clutter. Background subtraction is done for separating foreground objects from the background in a sequence of video frames. To detect the existence and direction of motion in a frame, a template matching approach called Motion History Image (MHI) is used. Instead of taking the Motion History Image for all the frames from start to end, the Motion History Image is taken only for selected key frame both before and after the interaction to improve the recognition rate. For accurate detection of action, Local Binary Pattern (LBP) and Histogram of Oriented Gradient (HOG) have been used. The similar actions are grouped together using k-means clustering. For training and testing, Support Vector Machine (SVM) has been employed as a classifier.*

Key Words: **Human Action Recognition, MHI, HOG, LBP, SVM.**

1. INTRODUCTION

Human Action Recognition (HAR) is an important and active field of research having a wide range of applications in numerous fields. HAR is widely employed in applications like police investigation camera, health care system and human-computer interaction. HAR system is projected to notice the action and to acknowledge the action as normal or abnormal action. The detection of human action recognition from a surveillance camera could be quiet complicated task because it holds more spatio-temporal information. The accuracy of recognizing the action is affected by viewpoint changes, background clutter and partial occlusion. The detection of body posture, gesture and key poses are the challenging task. Activity recognition is required for detecting activity in public places, daycare and healthcare monitoring system. Human actions usually consist of multiple persons. Acknowledgment of activity performed by various individual (e.g., punching, pushing, beating and caring) is difficult because of variations in motion, dynamic background, recording settings and interpersonal differences.

Previous research involves recognizing simple one person action like walking, jogging, hand clapping etc., [5] [6]. Practically thinking, these types of single person activities will not be able to occur in public. Human actions usually consist of multiple persons. Recognition of multiple person activities (e.g., punching, pushing, and handshaking) will be obligatory for several applications like, automatic detection of brutal activities in intellectual surveillance system [7]. The objective of this research is to find a strategy to detect two-person interactions efficiently. To detect the existence and direction of motion in a frame, a template matching approach named Motion History Image (MHI) is used. This technique has been used before in numerous cases for single person action recognition [1] [2]. In this paper, we have proposed an approach to recognize the actual frame where the interaction between two persons takes place. Our proposal introduces a modified version of MHI since basic MHI fails to generate appreciable outcomes [2]. Once the frame of interaction is detected, instead of taking the MHI of all the frames from start to end, the MHI is taken only for the key frame. The action detected frame in a video is called as key frame. For accurate detection of action, Local Binary Pattern (LBP) and Histogram of Oriented Gradient (HOG) have been used. The label for each action is generated using K-means algorithm. For training and testing, Support Vector Machine (SVM) has been employed as classifier. This paper is described as follows: Section 2 describes our proposed method. Experimental result and analysis are covered in section 3. Section 4 concludes the paper with a few future research scopes.

2. METHOD:

The Figure 1 illustrates the overall process for Human Activity Recognition (HAR) using the spatio-temporal feature. The input for the system is video. The first process is pre-processing of video. From the input video, Frames are extracted. The key frames are extracted by frame

differencing method. The current frame is subtracted from previous frame. The frame value other than zero is action detected frame. The action detected frame is key frame. The background subtraction is done only for the extracted key frames in order to improve the action recognition. The existence and direction of motion in a frame is identified by using MHI. The MHI contains both spatial and temporal feature. Then the local binary pattern and HOG feature is extracted. The HOG value gives the direction and magnitude information. The LBP feature used to find the texture. The k-means clustering is used to cluster similar groups with its centroid value. The grouping is done based on the extracted feature value. Based on the centroid value for each group label is given for each action. The label is given as input for classification using SVM. Then each activity is classified using SVM classifier and detects whether the action is normal or abnormal action.

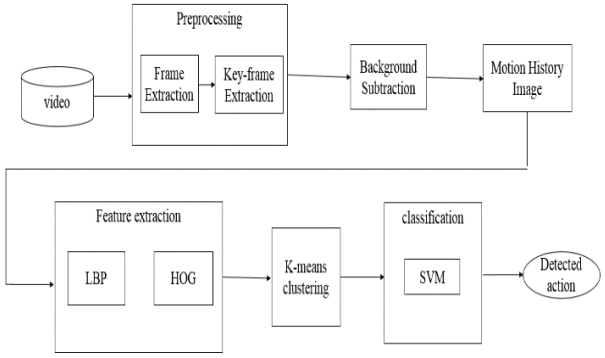

Figure 1: System Architecture for HAR using Spatio-Temporal Feature.

2.1 DATASET COLLECTION

The dataset is collected from UT-Interaction dataset and recorded for four different actions. These are:

1. Beating.
2. Caring.
3. Pushing.
4. Punching.

The dataset consist of two persons and are in outdoor and indoor environment. The dataset are taken in different background. There are 100 training videos with different action and 50 test videos. Each action is carried out by different persons in different videos. As a result, the inevitable challenge of variations in colors is encountered. The Figure 2 shows some sample frames of UT interaction dataset. Each action is carried out by different persons in different videos. The actions include beating, punching, pushing and caring.

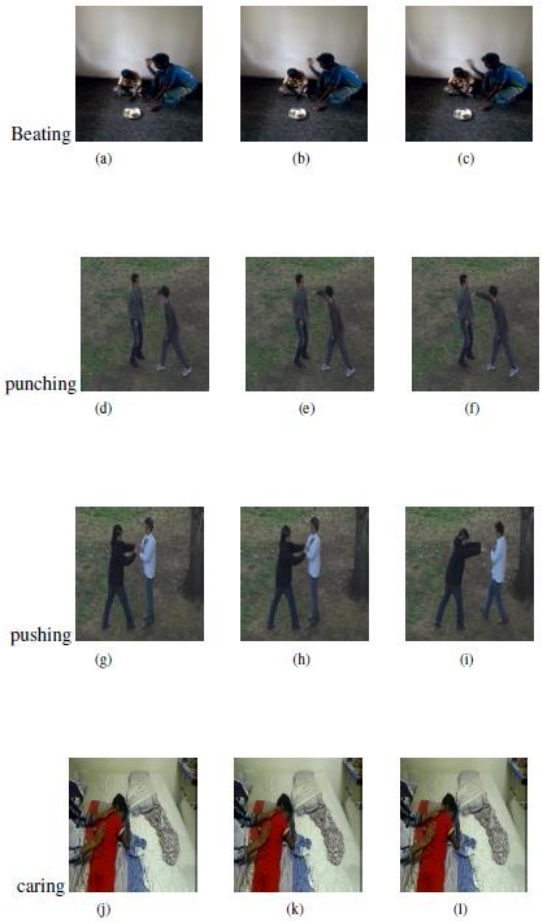

Figure 2: Sample Dataset for Different Action

2.2 BACKGROUND SUBTRACTION

Foreground extraction also called as background subtraction. Background subtraction is the process of separating the foreground objects from the background in a sequence of video frames. Background Subtraction generates a foreground mask for every frame. This step is performed by subtracting the background image from the current frame. When the background view excludes the foreground objects, it becomes obvious that the foreground objects can be obtained by comparing the background image with the current video frame. By applying this approach to each frame, the tracking of any moving object can be done. Background subtraction methods are widely used for moving object detection in videos in many applications, such as traffic monitoring, human motion capture and video surveillance.

The background subtraction is done to improve the recognition rate. The moving regions are detected by subtracting the current image pixel-by-pixel from a

background image that is created by averaging images over time in an initialization period. The basic idea of background subtraction method is to initialize a background firstly and then subtracting current frame in which the moving object present in that current frame is subtracted with background frame to detect moving object. This method is simple and accurately extracts the characteristics of target data, but it is sensitive to the change of external environment.

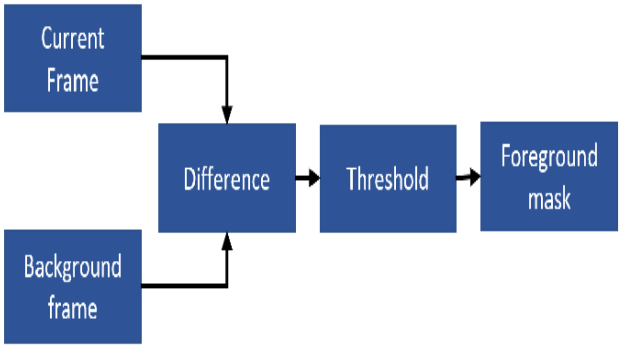

Figure 3: Block Diagram for Background Subtraction

The Figure 3 explains the background subtraction method. The first process is to initialize a background first and then subtracting current frame with background frame. Threshold is set for the differenced frame. Based on that threshold value the foreground is extracted. The foreground extraction process helps to identify the human action more accurately.

2.3 MOTION HISTORY IMAGE

The existence and direction of motion in a frame is identified by using MHI. MHI is a temporal template matching approach. Template matching approaches are the summation of immediate past successive images and the weighted intensity decays as time elapses. MHI is a cumulative grayscale images formed by spatio-temporal motion information. MHI expresses the motion flow of a video sequence in a temporal manner. MHI contains motion information and the direction of the motion. In MHI image, older movement portions in a video becomes darker than newer recent movement or moving regions. By subtracting two neighbouring and consecutive images the moving object can be detected. Then this image is converted into binary image. By layering the successive binary images, the Motion History Image are produced. The MHI are generated using frame differencing method. In an MHI image, more recently moving pixels are brighter and the image obtained is a scalar-valued image.

Figure 4: Sample Frame for Pushing Action.

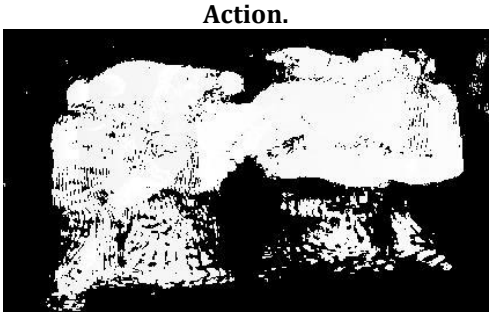

Figure 5: MHI Image for Pushing Action.

2.4 LOCAL BINARY PATTERN

The feature extraction is used to reduce the dimension of the action space by transforming it into feature representation. Features may be symbolic, numerical or both. An example of a symbolic feature is color and example of the numerical feature is weight. Features may also result from applying a feature extraction algorithm or operator to the input data. Local binary pattern texture operator converts the image into an array or an image of integer labels that describe small level changes in the image. A 3 x 3 neighborhood is formed around every pixel. Each pixel is subtracted with the center pixel value. If the result is a negative number then it is encoded as 0, otherwise 1. Then all these binary codes are concatenated in a clockwise direction to form a binary number. These derive binary numbers are called Local Binary Pattern.

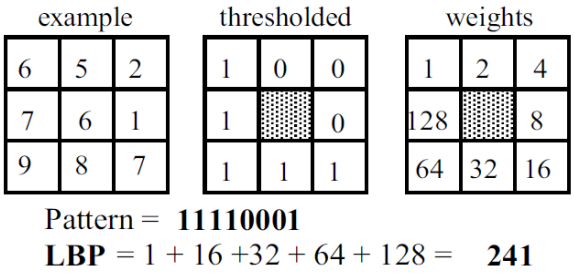

Figure 6: Example for Local Binary Pattern

2.5 HISTOGRAM OF ORIENTED GRADIENTS

HOG is a feature descriptor. The HOG is used to find the both directional and moving information in an image. It counts the number of occurrences of gradient orientation in the small patch of an image. In computing HOG at first, the image is divided into small connected regions called cells. Then a histogram of gradients within the cell is computed. A set of block histograms represents the descriptor. The HOG features contains both space and direction information. The HOG feature is calculated by the orientation of edge intensity gradients. The HOG feature is used to find both the shape of the object and direction information of the edge. The sobel filter is used to calculate the gradients dx(x,y) and dy(x,y) in x and y direction. By using this directional gradients, the magnitude M(x, y) and orientation (x, y) can be defined as

$$M(x,\ y) = \sqrt{dx(x,y)^2 + dy(x,y)^2} \quad (1)$$

$$\theta(x,\ y) = \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) \quad (2)$$

The Equation 1 is used to find the magnitude value and Equation 2 is used to find the orientation. The directional information explains the direction the action takes place. The magnitude information explains the force the action takes place. The image is resized as 64x128. The 64 (8x8) gradient vectors are generated. The generated gradient vectors are then represented as histograms. Each cell is split into angular bin. For, example 9 bins (0-180) 20 bins each. This splitting effectively reduce 64 vectors to just 9 values. These 9 values stores as gradient magnitude. Normalization is done in order to remove illumination changes. Finally block normalization is done.

2.6 CLUSTERING

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. A clustering algorithm is used to process feature vectors constituting the activity, by grouping them into clusters. The well-known k-means clustering algorithm, based on the squared Euclidean distance as a metric, can be used to group together the frames representing similar postures. The feature vector is loaded and k-means is called with the desired number of cluster. The distance from each centroid to points on a grid is computed. Then the cluster region is plotted.

2.7 CLASSIFICATION

SVM is a supervised learning method for classification and regression. 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. The advantage of k-fold cross validation is used for all the examples in the dataset are eventually used for both training and testing.

3. RESULT AND ANALYSIS

The Matlab is used for analysing the various human action. The video dataset is loaded first. From the given video frames are extracted. For further process only key frame is used. Background subtraction is done for all the key frames. The Motion History Image is identified from the background subtracted frame. The LBP and HOG feature value is extracted. Then, the extracted feature value is used for classifying the action as normal or abnormal.

The analysis is done for HAR using different algorithm. The chart-1 gives the accuracy for proposed system for different algorithm. The proposed system is tested on various algorithm with keyframe and without keyframe. The chart-2 explains the accuracy of different algorithm like SVM, K-Nearest Neighbour (KNN) and Ensemble algorithm. From the chart-2 it is proved that the accuracy of SVM is better when compared to other two algorithm. The accuracy of the system is further improved by considering keyframes. The action detected frame in a video is called as keyframe.

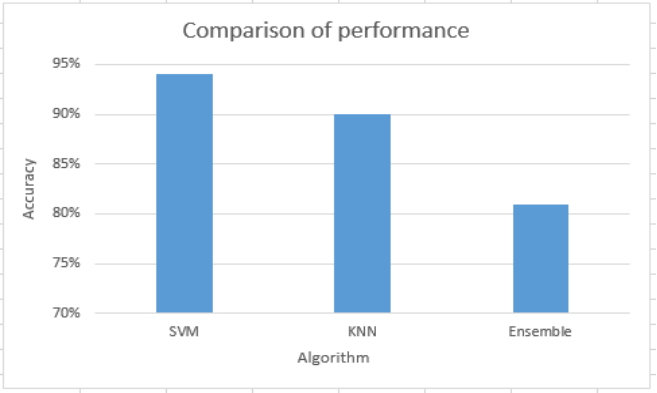

Chart-1: Accuracy Comparison for HAR using Supervised Algorithm.

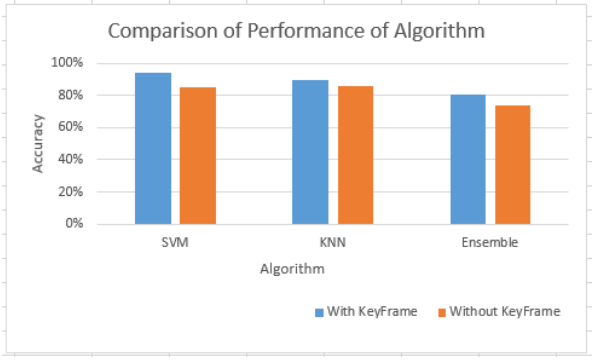

Chart-2: Accuracy Comparison for HAR using Supervised Algorithm with and without KeyFrame Extraction.

Chart-3 shows the comparison of different algorithm based on feature selection. The comparison is done by taking HOG feature and HOG combined with LBP feature. It is also proved that the accuracy of the System is improved by selecting both the HOG and LBP feature.

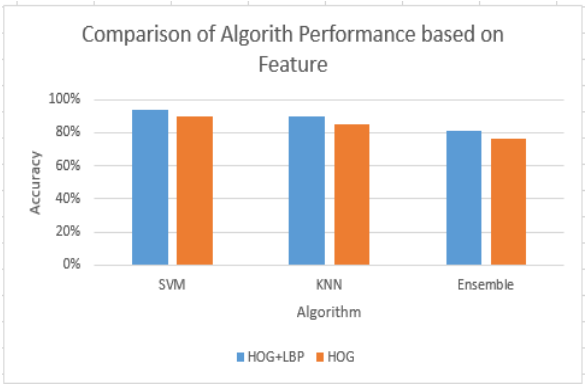

Chart-3: Accuracy Comparison for HAR using Supervised Algorithm with and without LBP Feature.

The system is tested on various supervised algorithm. The proposed system gets an accuracy of 90% on using linear SVM and 87% on using quadratic SVM. The system gives an accuracy of 85% by implementing KNN and 80% by implementing ensemble algorithm. The accuracy of the system is improved by extracting keyframes from the video. The accuracy is improved by 2% taking keyframe.

4. CONCLUSION AND FUTURE WORK

The video is converted into sequence of input frames. From the input frames the background is subtracted using multi-frame averaging method. The HOG and LBP features are obtained and then fused together. From the fused frame the centroids are obtained using k-means algorithm. SVM is used for classification. Each action is trained using SVM. The new data is tested on the proposed model. An evaluation of proposed work is performed on various data set and also an action is recognized from the new input data. Then the proposed system is compared with keyframe and without keyframe. The proposed system is also verified using different supervised algorithm. The performance of the system is evaluated using the confusion matrix and sensitivity which is formulated with True Positive Rate. The True Positive rate depicts the number of instances of a particular activity classified correctly out of the total number of instances of that specific activity. The confusion matrix help the user in better understanding of the accuracy of the classifier. The results shows that the proposed system classify the activity correctly and obtained the accuracy of 90%.

In this proposed work human action recognition is done using spatio temporal feature. The future work can be extended to other feature descriptor to reduce the false positive rate, so that the system can handle more complex videos and to improve the performance.

REFERENCES

[1] Alexandros Andre Chaaraoui and Jose Ramon Padilla-Lopez. “Evolutionary Joint Selection to Improve Human Action Recognition with RGB-D Devices”. In the Transactions of Journal of Expert system with applications, Vol. 41, pp 786-794, 2014.

[2] Alessandro Manzi, Flippo Cavallo and Paolo Dario. “A 3D Human Posture Approach for Activity Recognition Based on Depth Camera”. In the Proceedings of European Conference on computer vision, Vol. 9914, pp 432-447, 2016.

[3] Bruce Xiaohan, Caiming Xiong and Song-Chun Zhu. “Joint Action Recognition and Pose Estimation from Video”.

In the Proceedings of IEEE conference on computer vision and pattern recognition, pp 1293-1301, 2015.

[4] Haiam A, Abdul-Azim and Elsayed E.Hemayed. "Human Action Recognition using Trajectory-Based Representations". In the Transactions of Egyptian Informatics Journal, Vol. 16, pp 187-198, 2015.

[5] Jie Yang, Jian Cheng and Hanqing Lu. "Human Activity Recognition based on the Blob Features". In the Proceedings of IEEE International Conference on Multimedia and Expo, pp 358-361, 2009.

[6] K. G. Manosha Chathuramali, Sameera Ramasinghe and Ranga Rodrigo. "Abnormal Activity Recognition Using Spatio-Temporal Features". In the Proceedings of International Conference on Information and Automation for Sustainability, Vol. 39, pp 1–5, 2017.

[7] Kiwon Yun and Jean Honorio. "Two-person Interaction Detection Using Body-Pose Features and Multiple Instance Learning". In the Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp 28-35, 2012.

[8] Li Yao, Yunjian Liu and Shihui Huang. "Spatio-temporal Information for Human Action Recognition". In the Transactions of EURASIP Journal on Image and Video Processing, Vol. 39, pp 1-9, 2016.

[9] Maheshkumar Kolekar and Deba Prasad Dash. "Hidden Markov Model Based Human Activity Recognition using Shape and Optical flow Based Features". In the Proceedings of IEEE Region 10 Conference (TENCON), pp 393-397, 2017.

[10] Sheng Yu, Yun Cheng,Songzhi Su and Guorong Cai. "Stratified Pooling based Deep Convolutional Neural Networks for Human Action Recognition". In the Transactions of Multimedia Tools and Applications, Springer,, Vol. 76, pp 13367-13382, 2016.

[11] Vili Kellokumpu, Matti Pietikinen and Janne Heikkil. "Human Activity Recognition using Sequence of Posture". In the Proceedings of Conference on mission vision application, pp 570-573, 2015.

[12] Xin Yuan and Xubo Yang. "A Robust Human Action Recognition System using Single Camera". In the Proceedings of IEEE International Conference on Computational Intelligence and Software Engineering, pp 1-4, 2009.

BIOGRAPHIES

P.Sangeetha Received M.Tech and B.Tech in Anna University CEG campus, Guindy.