Comparative Analysis on Human Action Recognition Using Spatio Temporal Feature

G. Augusta Kani, Research Scholar, Anna University, Chennai & India, augus.jesus@gmail.com

P. Geetha, Assistant Professor, Anna University, Chennai & India, geethap@annauniv.edu

Abstract: **Human Interaction Detection is more essential in the arena of computer vision because of upward demands in several applications such as surveillance monitoring, entertainment and healthcare monitoring. Therefore, many research attempts have been undergo to precisely detect the human activities using data mining technique. The action recognition process involves extraction of information (features) from a video and to detect the interaction between two person using classifier. The recognition rate is affected by viewpoint deviations, illumination, partial occlusion, intersecting of two objects etc. From the sequence of video frames the motion is detected using subtraction method. Template matching approach is used to notice the occurrence of object and direction of motion in a frame named Motion History Image (MHI). To expand the recognition rate the MHI is taken only for particular key frame from both earlier and late human interaction instead of taking the Motion History Image from all the frames. Here Local Binary Pattern (LBP) and LBP based Histogram of Oriented Gradient (HOG) have been used for accurate discovery of action. The related actions are band together using dynamic k-means clustering. Real time and UT-Interaction dataset were used for training and testing process. Support Vector Machine (SVM) has been employed as a classifier which is compared with other classifiers such as KNN, Ensemble algorithm (EA) and proved that SVM approach delivers better accuracy to detect the actions exist in the videos.**

Keywords —Foreground Detection, Human Activity, Motion Detection, LBP, SVM, HOG, KNN, EA.

I. INTRODUCTION

Human Interaction Recognition (HIR) is a noteworthy and vigorous field of research employed in applications like police enquiry, human strength monitoring and human-computer interaction. Human Interaction (normal/abnormal activity) recognition from a surveillance camera may perhaps be quiet problematical job as it grasps more spatio-temporal information. Human actions usually consist of multiple persons, static and dynamic objects. Acknowledgment of activity performed by various individual (e.g., every single person consumes own way of gesture) is tough for the reason that of active background, inter, intra person transformations, recording sceneries, overlap objects and distinctions in motion. In this work the methodology en route for clip a pattern to discover two-person interactions proficiently and also it solve the problem of localization where the interaction takes place. This paper is planned as follows: Section II & III clarify the work related to our paper and defines our proposed method. Section IV & V displays the implementation of our approach and concludes the paper with a few forthcoming research scope.

II. RELATED WORK

A method combined with chromaticity and gradient to solve a shadow problem in background subtraction has been proposed by Xin Yuan and Xubo Yang [12]. A two-layered background subtraction is used which is based on both chromaticity and gradient to extract human contours from the frame sequence captured by the camera. This subtraction helps us to remove shadows from the foreground and get a good contour for recognition. The accuracy need to be improved in the system. Adaptive background subtraction algorithm has been implemented to obtain global outline feature and optical flow model to extract local visual feature by Jie Yang et al. [5]. Then these feature vectors are combined to form a hybrid feature vector. An action is recognized and accuracy is improved. The system cannot handle self-occluded object. Background subtraction is a technique for generating a foreground mask. For background subtraction, Sheng yu et al. [10] used to calculate the foreground mask from a subtraction between the current frame and background model. This technique consumes more time.

Shape can be represented by the boundary, region and moment. Canny edge detection technique is used to detect the edges of the image. The shape descriptor is used to describe the image content. Maheshkumar Kolekar and Deba Prasad Dash [9] have used canny edge detection for shape based feature extraction. Auto detection is inefficient. The shape-based methods capture the local feature from the human image. Alexandros Andre Chaaraoui and Jose Ramon Padilla [1] have obtained a silhouettes image from a frame using the silhouettes and optical flow based feature, and dense trajectory based feature. Computation cost is high for the system and early convergence is occurred. Human activity recognition in videos is important for content-based videos indexing. Li Yao et al. [8] proposed spatio-temporal

DOI : 10.35291/2454-9150.2019.0275 © 2019, IJREAM All Rights Reserved.

bio-graph based multi-feature fusion algorithm for detecting human action from spatiotemporal feature. The time consumption is more for large dataset. Kiwon Yun and Jean Honorio [7] have proposed multiple instance learning methods for detecting the interaction between two persons. Further, the classification of action is done using SVM. Dataset with multiple view point cannot be handled by this method. Action recognition and pose estimation from the video are closely related tasks for understanding human motion. Bruce Xiaohan et al. [3] proposed the and-Or graph model for extracting spatiotemporal feature. Manosha Chathuramali et al. [6] defined spatial information. The spatiotemporal feature contains both space and directional information. Illumination changes cannot be handled.

Feature extraction plays a vital role in the recognition of human action. Vili Kellokumpu et al. [11] proposed an SVM classifier for classifying the extracted feature to detect human action. The accuracy is less. SIFT descriptor is used to extract the feature and then classify the action. Haiam et al. [4] proposed a SIFT descriptor method for recognizing human action. By using the SIFT descriptor the recognizing rate can be increased. The recognition of the system needs to be improved. Alessandro Manzi et al. [2] used SVM classification for recognizing the activity of a single person. The SVM can candle more number of features for classification. Manosha Chathuramali et al. [6] used the SVM classification for classifying spatiotemporal feature to recognize only one person action.

III. PROPOSED METHOD

The complete procedure for Human Interaction Recognition (HIR) using the spatial-temporal feature is exemplified in Figure 1. The input for the structure is video. The first process is pre-processing of video. From the activity video, Frames are takeaway and stored in a folder (Figure 2). The key frames are extracted by frame differencing method which give more informative about the activity. The recent frame is subtracted from earlier frame. The frame value other than zero is action noticed frame. Due to reduce the time complexity background subtraction carried out only in a key frame. The action identified frame is considered as a key frame. MHI is used to locate the activity and their interaction which focus direction of activity also. The MHI contains both spatial and temporal feature. Then the local binary pattern and LBP based HOG feature is extracted. The HOG value gives the direction and magnitude data. The LBP feature used to catch the texture information. The dynamic k-means clustering is used to band related groups with its centroid value. The combination 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, KNN, and EA classifiers and detects whether the action is normal or abnormal action. The comparative analysis is done with other classifiers with and without key frames using spatio-temporal features.

A. BACKGROUND *SUBTRACTION*

Forefront extraction also called as background subtraction. Background subtraction is the procedure of splitting the forefront objects from the background in a sequence of video frames. Background Subtraction produces a foreground mask for every single frame, shown in Figure 3.

Algorithm 1. Background Subtraction

```
Input: Sequence of frame.
Output: Background subtracted frame.
1: Sets image dimensions as zero for all the input frames.
2: Accept and read the frames as img.
3: The average is obtained using mean from all the input frames and stored in bg-img
4: Then the img frame is doubled.
5: The img frame is subtracted with bg-img and stored in sub-img.
6: The result average background subtracted frame is shown and saved in folder.
```

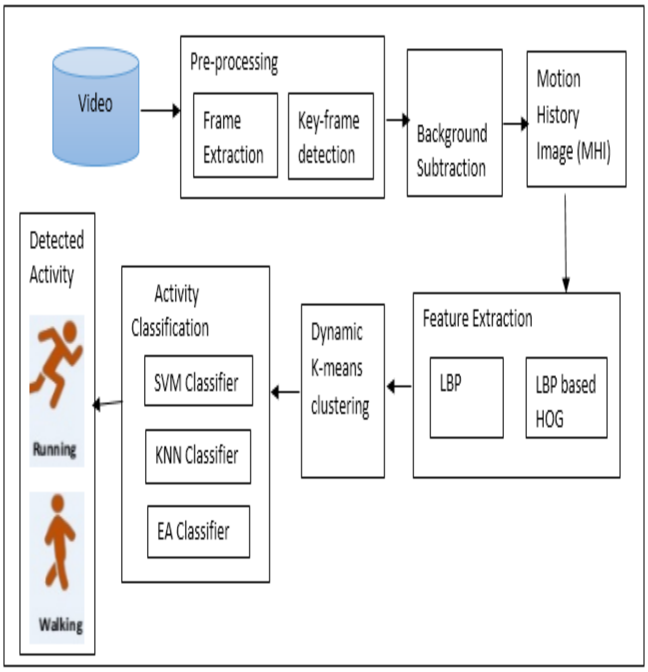


Figure 1: System Architecture for HIR using Spatio-Temporal Feature

This step is performed by subtracting the background image from the recent frame. When the background view excludes the forefront objects, it becomes obvious that the forefront objects can be obtained by comparing the background image with the recent video frame. By applying this approach (Figure 3) to each frame, the tracing of any motion can be done.

Figure 2: Extracted Key Frame from a Video

Algorithm 1 explain steps for background subtraction. Background subtraction methods are widely used for motion detection in videos in many applications, such as traffic monitoring, human motion capture and surveillance video.

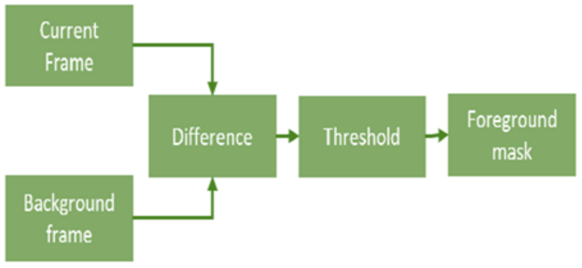


Figure 3: Block Diagram for Foreground Detection

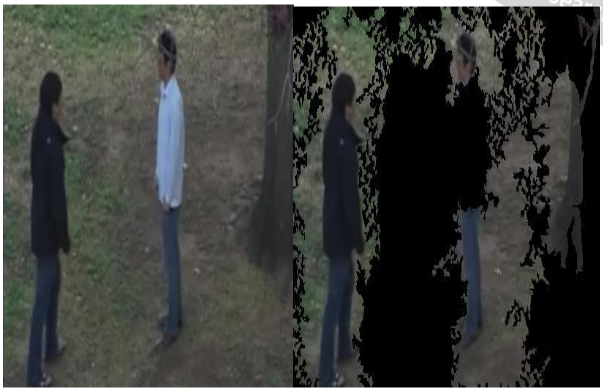


Figure 4: Input Frame and Background Subtracted Frame

B. FOREGROUND MOTION HISTORY IMAGE

MHI is a motion identification method used to locate the occurrence of object and their direction. MHI is a temporal template matching approach. Template matching approaches are the summary of instant earlier successive images and the weighted intensity decays as time elapses. MHI is a accumulative grayscale images formed by spatio-temporal motion data. MHI states the motion flow of a video sequence in a temporal manner. In MHI image, earlier motion in a video becomes darker than newer recent motion or moving regions. Then this image is converted into binary image. Motion History Images are shaped by layering the continuous binary images. The MHI are generated using frame differencing method. In that, a newly moving pixels are brighter i.e., positive and the image gained is a scalar-valued image. The Figure 5 shows the MHI for the given input video. The MHI contains the motion information of a video.

C. LOCAL BINARY PATTERN

The feature extraction is used to reduce the dimension of the action space by altering it into feature demonstration. Features may be symbolic, numerical or both. An example of a symbolic feature is color and example of the numerical feature is weight. 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.

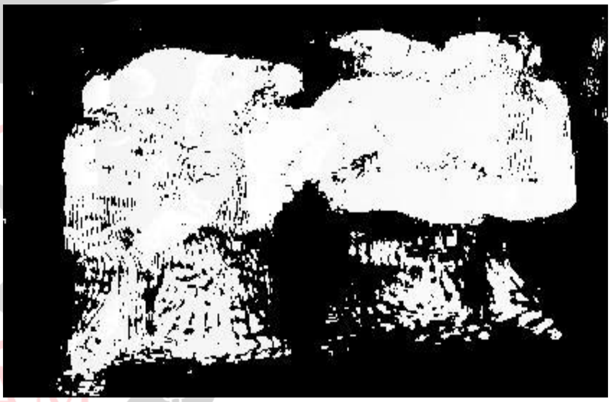


Figure 5: MHI Image for Pushing Action.

While the outcome is a less than zero then it is encoded as 0, or else 1. After that the concatenation is done in clockwise direction to form a binary number. These derive binary numbers are called Local Binary Pattern. Algorithm 2 explains the steps for finding LBP for an image.

Algorithm 2 Local Binary Pattern

```
Input: Motion History Image.
Output: LBP Code.
1: Read the image.
2: Form 3x3 cell.
3: If the result is positive then it is 1 else 0.
4: Concatenate in clockwise direction.
```

D. HISTOGRAM OF ORIENTED GRADIENTS

HOG is a feature descriptor. The HOG is used to find the both directional and moving information in an image. A set of block histograms represents the descriptor. The HOG features contains both space and direction information.

DOI : 10.35291/2454-9150.2019.0275
© 2019, IJREAM All Rights Reserved.

Algorithm 3 explains the steps for extracting HOG features. The HOG feature is calculated by the orientation of edge intensity gradients. 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) are given in Eq (1) and Eq (2).

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

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

The 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 is considered. This splitting effectively reduces 64 vectors to just 9 values. These 9 values are stored as gradient magnitude. Normalization is done in order to remove illumination changes. Finally block normalization is done. The feature value of HOG is 1xN where, N depends on the size of the image. The N vaue is computed as per the Equation (3). The blocks of an image is calculated using Equation (4).

$$N = Blocks\ in\ Frame \times Block\ Size \times Bin\ number \quad (3)$$

$$Blocks\ in\ Frame = \left\{\frac{\frac{Frame\ Size}{Cell\ size} - Block\ size}{Block\ Size - Overlap} + 1\right\} \quad (4)$$

```
Algorithm 3 Histogram of Oriented Gradient
Input: Frames.
Output: HOG feature.
1: Read the image.
2: Sobel filter calculate the gradient in both direction dx(x,y) and dy(x,y).
3: By using gradients, the magnitude and orientation is calculated.
4: Normalization is done for HOG vector.
5: HOG feature length is calculated.
6: Finally the feature extracted over all the planes are combined together.
```

E. DYNAMIC K-MEANS CLUSTERING

Cluster analysis 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. Algorithm 4 explains the steps for Dynamic K-means clustering. It is used to process feature vectors constituting the activity, by grouping them into clusters. The eminent dynamic k-means based on the squared Euclidean distance as a metric, can be used to group together the frames representing similar postures. The cluster region is plotted according to the number of cluster. The steps for dynamic K-means clustering is given below. The Figure 6 shows grouping of four different actions like beating, Pushing, Punching and Caring with its centroid value. Based on the centroid value labels are given for each action.

```
Algorithm 4 Dynamic K-means Clustering
Input: Feature vector.
Output: Centroid.
1: Load the feature vectors.
2: Call k-means with k, the desired number of clusters.
3: Compute the distance from each centroid to points on a grid.
4: Plot the cluster region.
```

F. CLASSIFICATION

SVM is a supervised learning method for classification and regression. SVM is 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 inputs are the images of three categories 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. The output for abnormal video classification and email notification are shown in Figure 9. If the classification classifies the output as abnormal then an email alert is sent.

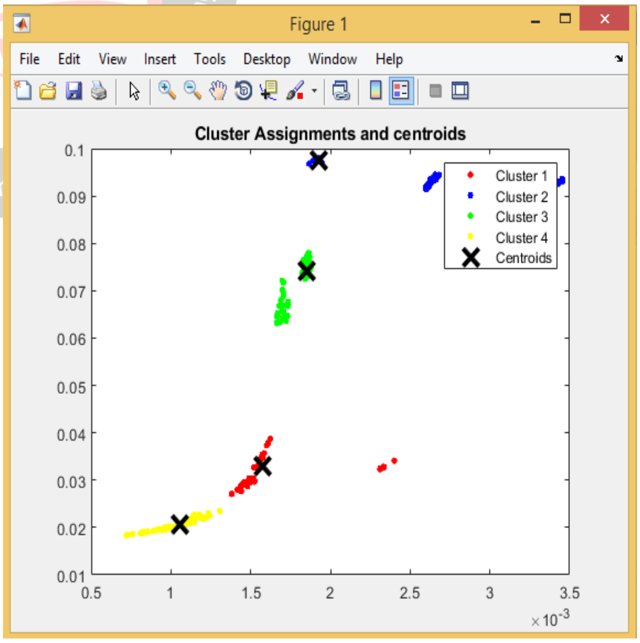


Figure 6: Dynamic clustering for the given Feature Value

IV. RESULT AND ANALYSIS

The data is collected from UT-Interaction dataset different normal (i.e walking, caring, handshaking) and abnormal actions (i.e pushing, punching, fighting, boxing). The dataset consist of two persons and are in outdoor and indoor

DOI : 10.35291/2454-9150.2019.0275
© 2019, IJREAM All Rights Reserved.

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 Matlab is used for analyzing the various human actions and it's concisely described in each algorithm steps. 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 HIR using different algorithm. The Figure 7 gives the accuracy for proposed system for different algorithm. The proposed system is tested on various algorithm with key frame and without key frame. Figure 8 explains the accuracy of different algorithm like SVM, K-Nearest Neighbour (KNN) and Ensemble algorithm. From the Figure 8 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 key frames. The action detected frame in a video is called as key frame.

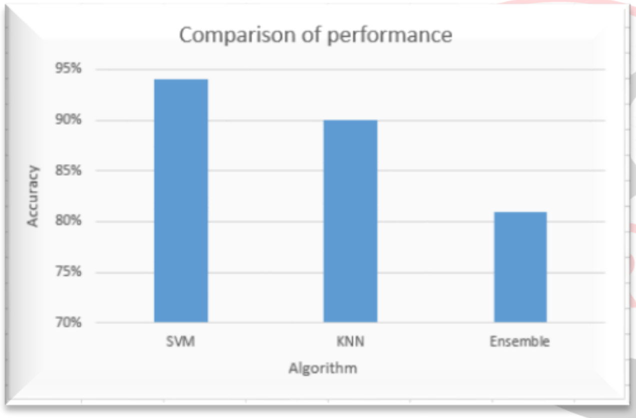


Figure 7: Accuracy Comparison for HIR using Various Supervised Algorithm.

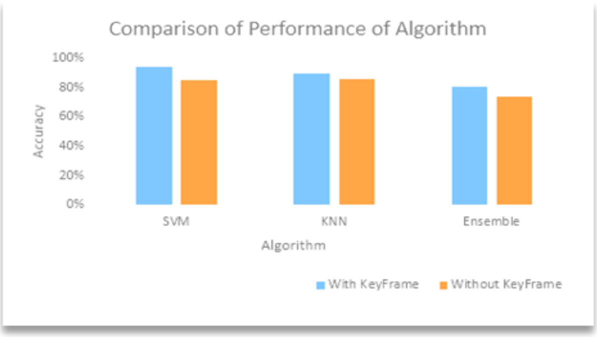
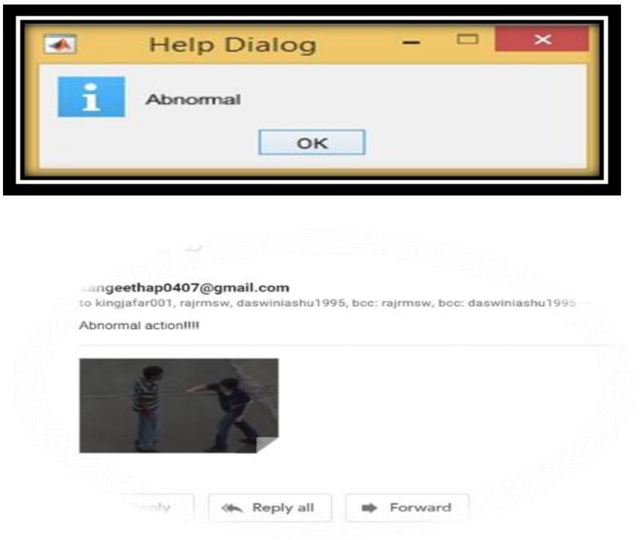


Figure 8: Accuracy Comparison for HIR using Supervised Algorithm with and without Key-Frames Extraction.



Help Dialog

Abnormal

OK

…angeethap0407@gmail.com
to kingjafar001, rajrmsw, daswiniashu1995, bcc: rajrmsw, bcc: daswiniashu1995

Abnormal action!!!!

Figure 9: Classification Output for Abnormal activity & E-mail Alert

Figure 10 shows the evaluation 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 upgraded by selecting both the HOG and LBP feature.

Activity Recognition	Accuracy (%)
Key Frame + LBP based HOG + Linear SVM	88
Key Frame + LBP based HOG + Quadratic SVM	85
KNN	82
ENSEMBLE	80

Table 1: Performance Analysis of Different Algorithm

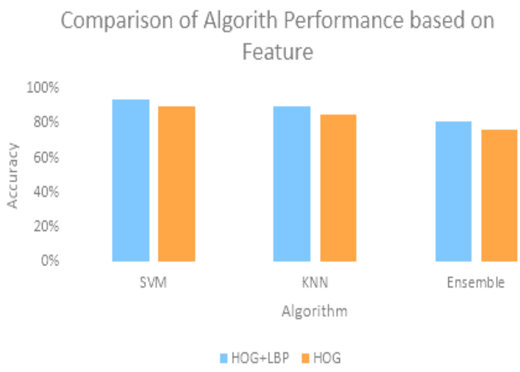


Figure 10: Accuracy Comparison for HIR using Supervised Algorithm with and without LBP Feature.

The system is tested on various supervised algorithm shown in Table 1. The proposed system becomes an accuracy of 88% on using linear SVM and 85% on using quadratic SVM. The system gives an accuracy of 82% by implementing KNN and 80% by implementing ensemble algorithm. The overall accuracy of the system is improved 2% by recognizing the actions from key-frames of video.

DOI : 10.35291/2454-9150.2019.0275 © 2019, IJREAM All Rights Reserved.

V. CONCLUSION AND FUTURE WORK

The proposed work used to verify different supervised algorithms (SVM, KN, EA) with certain combination of information to expand the recognition rate. Here video frames are transformed into series of key frames to reduce both the time and processing complexity. The LBP and LBP based HOG features are obtained and then fused together for giving rich information regarding object presence. From the fused frame the centroid is obtained using dynamic k-means algorithm. Each action is trained using SVM and other classifiers. 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. The performance is tested by with key-frame and without key-frame. The results were shown that the comparison of various supervised algorithm for the action recognition from the given video data set is implemented, analyzed and SVM obtained the high accuracy of 88%.

In this work, human action interaction is done using spatio temporal feature. The forthcoming work can be prolonged to other feature descriptor to reduce the false positive rate, so that the system can handle more complex videos and 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.

DOI : 10.35291/2454-9150.2019.0275 © 2019, IJREAM All Rights Reserved.