Geometric Invariant Model Based Human Action Recognition

Pravin Nagar
Information Technology
Indian Institute of Information Technology
Allahabad, India
pravinnagar619@gmail.com

Anupam Agrawal
Information Technology
Indian Institute of Information Technology
Allahabad, India
anupam69@gmail.com

***Abstract*—Most of the state-of-the-art methods for action recognition are very complex and variant to the geometric transformation like scaling, translation and rotation. Cuboid based method required all frames to extract the cuboid of action that's why cuboid based methods are expensive. Other methods use contour based approach for feature representation which is not robust to noise. So we require a very fast and robust feature descriptor which is invariant to geometric transformations. To deal with the above challenges our approach employs a geometric invariant model based human action recognition. It uses R-transform for feature representation. From each video we require a limited (approx. 10-15) number of frames and after detecting normalized foreground, we apply R-transform on Reason of Interest. The features of R-transform are: it is invariant to RST (rotation, scaling and translation), robust to noise and its complexity is NlogN where N=size of image i.e. N=n*n. When we are using PCA and LDA for dimension reduction and ANN (Artificial Neural Network) for classification the accuracy of our method falls in between 90 to 96% and with the PCA and Euclidian Distance based Classifier it falls in between 87 to 92%.**

Keywords— ***geometric invariant; silhouette based; R-transform; model based.***

I. INTRODUCTION

The analysis of human action is an active research area of computer vision because of the wide range of vision related applications such as surveillance based application, video categorization and tagging, context-aware pervasive systems, gait recognition and human-computer interaction based applications. And if we compare machine's classification power with human being's classification power, despite of enough technological advancement the machine action classification has many challenges like confined to limited number of action classes, not robust to geometric transformation like rotation, scaling and translation, computationally very costly and suffers from dynamic and cluttered background, abrupt illumination, occlusions etc. so we need a system which is immune to all of these challenges. For any method as soon as you increase the number of active class accuracy decreases Hence human action recognition (HAR) is still a very challenging problem in computer vision.

Our emphasis is to design a fast method which doesn't require all frames of video sequence to process and also invariant to geometric transformations. These two things ensure the features from camera orientation and mimic the human tendency to not process all frames for action recognition. In human action recognition domain the robustness of the method to the above challenges is ensured by the feature extraction process. In this work we are using a constant number of frames for feature representation. After foreground detection we are taking Radon transform of a human silhouette. In surveillance based application background is usually very complex and shape of human also varies from its distance to the camera. Because Radon transform is variant to the geometric transformation that's why we are taking R-transform for each human silhouette so that our feature is robust to noise and invariant to geometric transformations. R-transform which is an adaptation of Radon transform is a new feature descriptor which is robust to holes in binary silhouette, frame losses and effective to computation. Moreover, many experiments proved its superiority over other model based feature descriptors in human action recognition [1].

After getting the feature which is invariant to geometric transformation we are required to reduce the dimension of feature vector to speed up calculation and removing the redundancy so we are using PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) for dimension reduction. The output of this process gives you a powerful discriminant feature vector which is used for classification. For classification of testing video we use Euclidian distance based classifier (linear classifier) and ANN (non-linear classifier). This method is tested on a KTH dataset (6 action classes), Weizmann dataset (10 action classes) and In-house dataset (9 action classes).

II. RELATED WORK

Initially sensor based approach being used for human action recognition, but this approach is neither cost effective

nor subjected to a real life environment [2]. At a glance the whole human action recognition approach is divided into two parts, first is feature representation and the second is classification.

A. Feature Representation

At a broader level feature representation for HAR is divided into various classes and subclasses as shown in fig. 1. First is preserving the background and second is without preserving the background. Kishor K Reddy and Mubarak Shah [3] have proposed a very efficient method for a large class of action which falls in the first class of feature representation. In this method they used fusion of motion descriptor and scene context descriptor with the help of 3D SIFT descriptor. Motion descriptor keeps track the motion present in the sequence and scene descriptor keeps track the scene information, i.e. where the activity is being performed. The second class is without preserving the background that means background subtraction. Background subtraction is further divided into model based, model less and hybrid method [2].

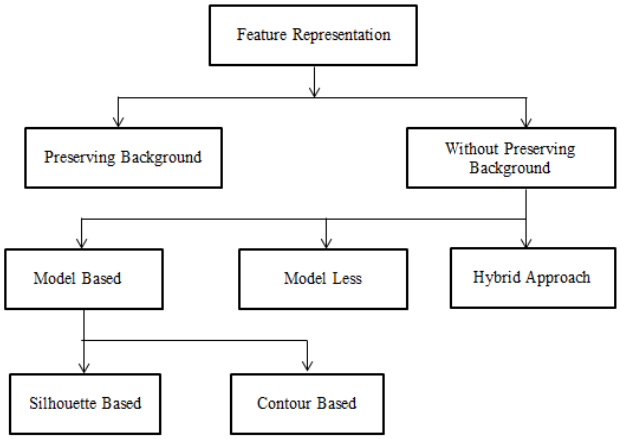

Fig . 1 various categories of feature representation

In Model Based approach features are extracted from the geometry of the human silhouette. The Model Based approach further divided into two parts, namely Silhouette Based and Contour Based. Silhouette Based methods consider all pixels in the feature extraction process, whereas Contour Based approach only takes pixels present on the boundary of the human shape. Many works of literature used Fourier Descriptor, Wavelet Descriptor and Hough transform for contour based method [4] [5]. This is also a reason why Contour Based method suffers from noises. It also doesn't account the internal geometry of human shape so it is also confined to limited applications [1]. Silhouette Based methods are a good low level feature described as comparable to Contour Based method.

The model Less approach doesn't concern with the geometry of the human body or we can say that while taking features the geometry of human body distorted. The features comes under model less approach are accumulated motion based features, gradient of motion, shape based features, energy images, histogram of images, etc. Model less approaches are robust to noise but have high computational complexity.

There is a trade-off between model less and model based approach. Hybrid approach is used to overcome the shortcoming of both the approaches [2].

B. Classification

In literature as such no best classifier is suggested because classifier is just a tool to classify the input features space and your result accuracy depends upon the discrimination power of your feature space [2]. But still according to the requirement of applications some classifier is superior then others like if we have limited training data set, then Support Vector Machine is a superior classifier than other classifiers. But we have sufficient training dataset so we are using Distance Based Classifier (Linear) and Artificial Neural Network (Non-linear) [2].

III. PROPOSED METHODOLOGY

There are two types of term in literature for human motion analysis being used namely action and activity. Both look similar but having minor differences action is defined as a motion pattern generated by a single person without intervention of any object, whereas activity is performed by either multi person or by a single person with some object [6]. The overall system is divided into three parts, namely pre-processing module, feature extraction and classification as shown in fig. 2.

A. Pre-processing

1. Frame extraction
2. Background Subtraction
3. Silhouette Detection and Normalization

The input video sequence could have different formats and quality. But our system doesn't require all frames for the process because it is wastage of computational resources. A typical input video is 30 fps, but we don't need to process all the frames, our system only required 15 uniformly distributed frames throughout the video so this overcomes the unnecessary processing. Many methods in the literature are bound to process all frames in the video to find the trajectory of interest points, but this is not the case with our method. For background subtraction we have used frame differencing in which the previously taken background frame is subtracted

from input video frame. So the output of this sub-module would be a 15 uniformly distributed foreground human blob. After that the human silhouette is detected from the foreground and enclosed is a rectangular box and this enclosed box which is also called as region of interest is normalized to 50*50 box for the feature extraction process. So ultimately we would have 15 regions of interest of normalized 50*50 size.

B. Feature Extraction and Dimensionality Reduction

For feature extraction a new shape descriptor which is an extension of Radon transform called R-transform is used. Radon transform is set of projection of ROI along a line which makes an angle θ with *Y* axis and perpendicularly ρ distance apart from the origin [7] [8]. The complexity of Radon transform is NlogN where N=n*n binary image. The mathematical representation of Radon transform is as follows:

$$T_{rf}(\rho,\theta) = \sum_x \sum_y f(x,y)\delta(x_0 \cos\theta - y_0 \sin\theta - \rho) \quad (1)$$

Where $f(x,y)$ is the binary image and define as a

$$f(x,y) = \begin{cases} 1 & if\ f(x,y) \in D \\ 0 & \text{otherwise} \end{cases}$$

Where D is a domain of binary image and $\in [0,\pi)$.

R-transform is an adaptation of Radon transforming because Radon transform is not invariant to geometric transform. R- Transform is defined as the sum of the square values of Radon transform for all the lines in an image having the same angle θ [9] [10].

The mathematical representation of R-transform is as follows

$$R_{f(\theta)} = \sum_\rho (T_{rf}(\rho,\theta))^2 \quad (2)$$

Where T_{rf} is Radon transform of $f(x,y)$.

R-transform is invariant to translation and scaling and gives phase shift on rotation. In mathematical form, we can represent these properties as follows [7]:

Periodicity: The period for the R-transform is π and mathematically it is represented by following equation.

$$R_f(\theta) = R_f(\theta \pm \pi) \quad (3)$$

Translation: R-transform is invariant to translation of binary image defined by f by some vector U= (x_0, y_0).

$$\sum_\rho {T_{rf}}^2 (x_0 \cos\theta - y_0 \sin\theta - \rho) = R_f(\theta) \quad (4)$$

Rotation: A rotation of the image by an angle θ_0 implies a translation of the R-transform of θ_0.

$$\sum_\rho {T_{rf}}^2 (\rho, \theta + \theta_0) = R_f(\theta + \theta_0) \quad (5)$$

Scaling: Scaling in f by some factor α produces a change in amplitude of the R-transform.

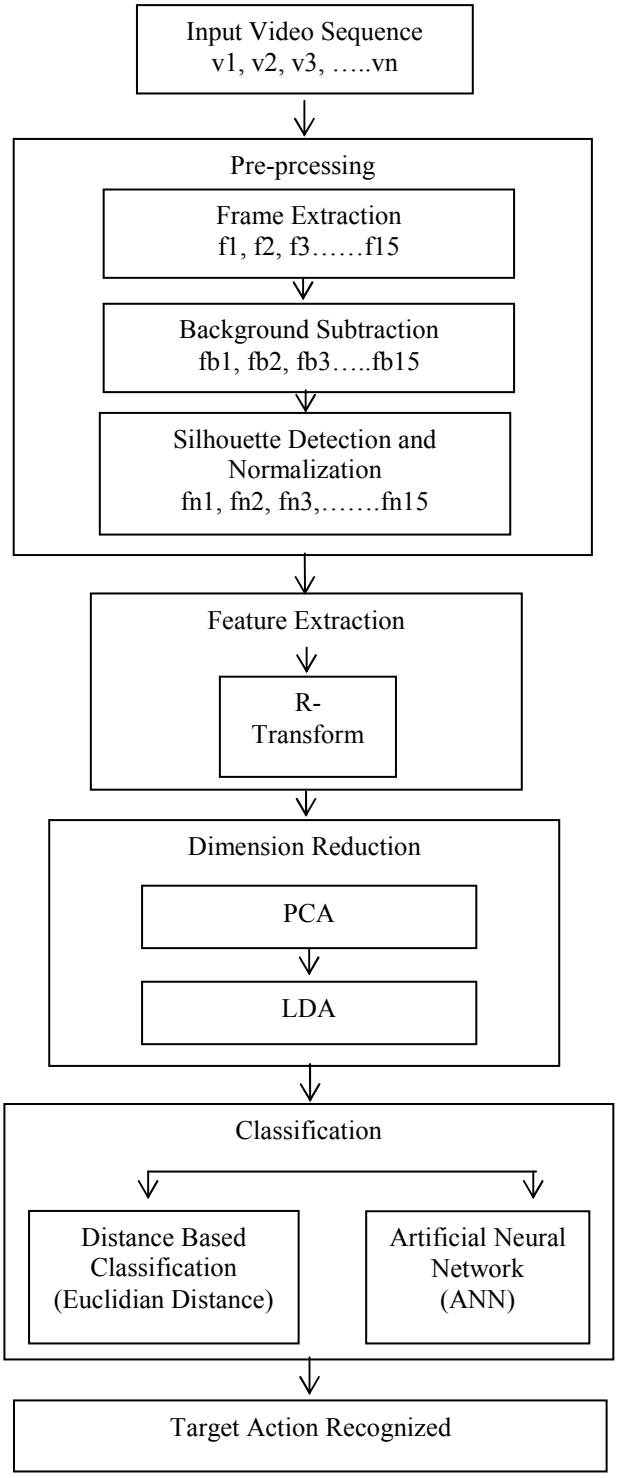

Fig.2 Flow chart of Proposed Methodology

In in this work we are taking Radon transform of region of interest for sixty angles, i.e. 3, 6, 9… 180 to calculate R-transform. So for each angle we get a column vector in which each row is a projection along one line. Now our binary n*n image is converted to 75*1 column vector. For each angle we get one 75*1 column vector and for 60 angles we get 75*60

matrix. After that to find R-transform squared sum all column vectors because R-transform is defined as the squared sum of all projection lines for same angle θ. Now we get a row vector of 1*60, take a transpose of that we get an R-transform column vector which represents one binary image which is geometric invariant and very less computational complex.

For dimension reduction we are using PCA and LDA separately and sequentially. After applying PCA and then LDA 60*15 matrix is reduced to 40*15 so computation is reduced considerably. The dimensionality reduction ensures faster processing of input video sequence.

A. Classification:

As we have discussed earlier that there is no best classifier in machine learning and some classifiers are good for specific application. So keeping in mind these things we are using Euclidian Distance based Classifier which is a linear and Artificial Neural Network a nonlinear classifier. The system is giving very good classification result and system accuracy is reached up to 95% above. We have tested our method on Weizmann dataset; out of 90 video only 6 are misclassified so approximately it is giving us 93% accuracy, using distance based classifier and above 95 % using nonlinear classifier.

IV. Experiments and Results

A. System Setup: Experiments was conducted on an Intel Core i3 second generation processor clocked at 2.4 GHz with 4GB of RAM. Processing was carried out at 160*120, 180*144 and 240*180 resolution of input frame of KTH, Weizmann and In-House datasets respectively. Average processing speed of around 1/6 fps was obtained for the Weizmann dataset and 1/9 fps for the In-House dataset. We have tested our method using a system stop watch and it can classify 20 videos in 36 seconds.

*B. Datasets***:**

1. *Action Recognition on KTH dataset:*

KTH is a standard dataset contains six action classes; we have taken total 60 video sequences, 10 video sequences for each class. The action classes are boxing, hand clapping, hand waving, jogging, running, walking with 160*120 resolution. We have taken 70% videos for training and remaining for testing and we are getting 90% accuracy with our full-fledged system and confusion table is not shown because of lack of space.

2. *Action Recognition on Weizmann dataset***:**

The Weizmann dataset is a standard dataset comprising 90 video sequences. There are total 10 action classes and each action is performed by 9 persons in the homogenous outdoor background with static camera and 180*144 resolution as shown in fig. 3. We have taken seven videos from each action class for training and remaining two for testing purpose. If we use PCA and LDA for dimension reduction and Euclidian distance based classifier for classification then we are getting average 92 % recognition rate. Our full-fledged system comprises R-transform for feature extraction, PCA and LDA sequentially for dimension reduction and ANN for classification with this system our average recognition rate is above 95% and the confusion table of Weizmann dataset using ANN based classifier is shown in Table 2.

Fig.3 Snapshot of Weizmann data set containing ten types of human action (Bend, Jumping Jack, Wave1, Wave2 Jump, Skip, Walk, Run, Pjump, Side,) and In-House data set containing nine types of human action (Bend, Jumping Jack, Wave1, Wave2 Jump, Skip, Walk, Run, Pjump, Side,)

3. *Action Recognition on In-House dataset***:**

We have tested our method on the In-House dataset containing 63 video sequences. Each action class contains 7 videos which is performed by 5 persons. All videos are taken by static camera and having 240*180 resolution as shown in fig 3. For training we are using 5 videos and remaining 2 are for testing purpose. If we use PCA and LDA for dimension reduction and Euclidian distance based classifier for classification then we are getting average 91 % recognition rate. Our full-fledged system comprises R-transform for feature extraction, PCA and LDA sequentially for dimension reduction and ANN (Artificial Neural Network) for classification with this system our average recognition rate is approx. 94%. The confusion table of In-House dataset with ANN based classification are shown in Table 3.

Table 1: Percentage Accuracy of Various Method

Datasets	KTH	Weizmann
Proposed Method	90	95
Cuboid Based Methods [12] [13]	93	96.7
Bregonzio et al. [11]	93.17	96.66
Dollar et al. [14]	81.17	85.2

Table 1 shows comparison of accuracy on KTH and Weizmann datasets by various methods. There are some methods whose accuracy are better than our method, but they are variant to basic geometric transformation and processing all frames to extract features.

C. Results Summarization:

1. Radon transform of two similar or confusing action classes: Fig. 4 show discriminating power of Radon transform with two confusion action classes of Weizmann dataset, where the X axis is the magnitude of binary silhouette and the Y axis is the distance of the projection line from the origin i.e. .

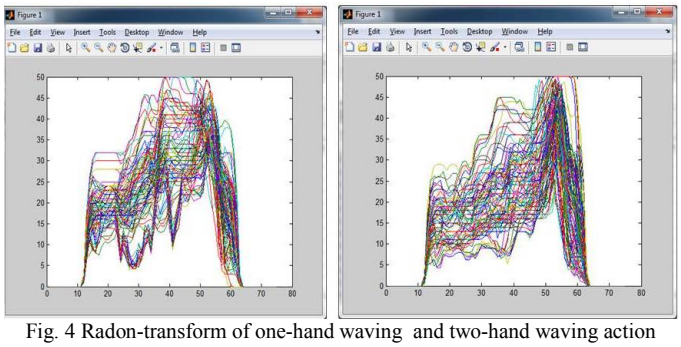

Fig. 4 Radon-transform of one-hand waving and two-hand waving action classes

2. Confusion Matrix of results on two datasets:

Table 2: Confusion table for Weizmann data set using Artificial Neural Network

Weizmann Dataset of 10 Action Classes										
Act	1	2	3	4	5	6	7	8	9	10
A1	**1**									
A2		**.89**		.11						
A3			**.89**							.11
A4				**.89**						.11
A5					**1**					
A6						**.89**		.11		
A7							**1**			
A8								**1**		
A9						.11			**.89**	
A10		.11								**.89**

3. Accuracy assessment of changing various parameters: We have conducted our experiments by changing various parameters and compare their effects on the results. Our first try comprises Radon transform for feature extraction, PCA and LDA sequentially for dimension reduction and distance based classifiers for classification. With this arrangement itself results are quite satisfactory and percentage accuracy is 92% on Weizmann dataset and approximately 91% on the In-House dataset. Instead of distance based classifier if we use artificial neural network, then average recognition rate further increases and accuracy reach of 95 %. But we have introduced the new feature descriptor called R-transform which is invariant to basic geometric transformation such as rotation, scaling and translation.

Table 3: Confusion table for In-House Dataset using Artificial Neural Network

In House dataset of 9 Action Classes									
Act	1	2	3	4	5	6	7	8	9
Bend	**1**								
JJ		**1**							
1 H			**1**						
2 H				**1**					
Jump					**1**				
Skip						**1**			
Walk						.145	**.71**		.145
Run								**.86**	.14
JSM						.14			**.86**

Table 4: Effects of various parameters on accuracy on Weizmann dataset

Features	**Dimension Reduction**	**Classifier**	**Percentage Accuracy**
Radon Transform	PCA+LDA	Distance based	88-92
Radon Transform	PCA+LDA	ANN	92-95
R-Transform	PCA & LDA	Distance based	87-92
R-Transform	PCA & LDA	ANN	92-96

If we apply R-transform for feature extraction, PCA and LDA sequentially for dimension reduction and distance based classifier is used, then our percentage accuracy in between 87-

90. Our full-fledged system comprises R-transform with sequentially PCA and LDA and non-linear classifier gives above 95% accuracy. Table 4 shows the response of accuracy corresponds to a particular parameter.

4. Comparison with the state-of-the-art Methods: There are various methods being used for human action recognition. [11] proposed a novel approach of video indexing which uses a combination of histogram of spatial sub-graph and temporal sub-graph generated by Space–Time Interest Points (STIP) features. The spatial sub - graph represents the intra frame information and temporal frame represents the inter frame information. [2] used internal and external features for action recognition where internal features represent internal information of blob and external features capture trajectory information; with this arrangement they achieved 90 percent accuracy on Weizmann dataset. [15] calculates DFT on accumulated motion image. [16] [17] and [18] detect the key points using linear separable filter and track them to find discriminant trajectory information. [19] proposed RMI feature extraction on the star skeletonization of the object to reduce the computational cost and method gives 96.87 percent accuracy but number of classes are only four.

So far we have discussed various approaches for action recognition but there are so many shortcoming and some of them are removed considerably in the proposed work. Like all the methods discussed above process all frames to extract features, especially trajectory based method, but our work just takes fifteen uniform frames from the input video sequence. Second is the features are not robust to basic geometric transformation which is the demand of various robotics based, surveillances based applications and virtual avatar.

V Conclusion and Future Scope

We have proposed an efficient feature extraction method called R-transform to represent each action class and which is invariant to geometric transform like translation, scaling and rotation, robust to noise and having very less computational complexity. We are using the binary image for calculation Radon transform of image blob so it would not capture the depth information so if we have a depth camera or gray image, then we can calculate Radon transform more efficiently and our features would be stronger and accuracy will also increase [20]. Apart from depth images if we increase the number of action classes we can fuse R-transform with BOVW (Bag of Visual Words) to generate stronger results [21].

References

[1] Ying Wang, Kaiqi Huang and Tieniu Tan, "Human Activity Recognition Based on R Transform", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2007.

[2] Htike, Z.Z., Egerton, S., and Kuang Ye Chow, "Real Time Human Activity Recognition Using External and Internal Spatial Features", IEEE Sixth International Conference on Intelligent Environments (IE), pp. 52 – 57, 2010.

[3] Kishor K. Reddy and Mubarak Shah "Recognizing 50 human action categories of web videos" Springer-Verlag international conference Machine Vision and Applications journal, Volume 24, Issue 5, pp. 971-981, 2012.

[4] Chuang, G.C.-H and Kuo, C.-C.J. "Wavelet Descriptor of Planar Curves: Theory and Applications", IEEE Transactions on Image Processing, Volume 5, Issue 1, pp. 56-70, 1996.

[5] Dengsheng Zhang, , Guojun Lu, , "Shape-based image retrieval using generic Fourier descriptor" Signal Processing: Image Communication Volume 17, pp. 825–848, 2002.

[6] Turaga, P., Chellappa, R., Subrahmanian, V.S. and Udrea, O. "Machine Recognition of Human Activities: A Survey", IEEE Transactions on Circuits and Systems for Video Technology, Volume 18, Issue 11, pp. 1473 – 1488, 2008.

[7] S. Tabbone, L. Wendling, and J.-P. Salmon, "A new shape descriptor defined on the Radon transform", Science Direct's Computer Vision and Image Understanding, Volume 102, pp. 42–51, April 2006.

[8] Zhijing Liu and Haiyong Zhao, "Shape-Based Human Activity Recognition Using Edit Distance", 2nd International Congress on Image and Signal Processing, pp. 1-4, 2009.

[9] Khan, Z.A. ; Won Sohn, "Feature extraction and dimensions reduction using R-transform and Principal Component Analysis for human activity recognition", IEEE Conference on Advanced Information Management and Service (IMS), pp. 253 – 258, 2010.

[10] Z.A. Khan and W. Sohn, "Hierarchical human activity recognition system based on R-transform and nonlinear kernel discriminant features", IET journals & magazines, pp. 1119 - 1120, 2012.

[11] M. Bregonzio, S. Gong and T. Xiang. "Recognising Action as Clouds of Space-Time Interest Points," IEEE International Conference on CVPR, pp. 1948–1955, 2009.

[12] Vishwakarma S., Agrawal, A, "Framework for human action recognition using spatial temporal based cuboids", International Conference on Image Information Processing (ICIIP), pp. 1-6, 2011.

[13] Vishwakarma S., Sapre A. and Agrawal, A., "Action recognition using cuboids of interest points", IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), pp. 1-6, 2011.

[14] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie., "Behavior Recognition via Sparse Spatio-Temporal Features", 2nd Joint IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, vol. 02, pp. 65–72, 2005.

[15] Thanikachalam, V. and Thyagharajan, K.K., "Human action recognition using accumulated motion and gradient of motion from video", IEEE International Conference on Computing Communication & Networking Technologies (ICCCNT), pp. 1 – 6, 2012.

[16] Messing, R., Pal, Chris and Kautz, H. "Activity recognition using the velocity histories of tracked keypoints", IEEE 12th International Conference on Computer Vision, pp. 104-111, 2009.

[17] A. Oikonomopoulos, I. Patras and M. Pantic, "An Implicit Spatiotemporal Shape Model for Human Activity Localization and Recognition", IEEE International Conference on CVPR, pp. 27-33, 2009.

[18] Fei Yuan, Gui-Song Xia, Sahbi, H. and Prinet, V. , "Spatio-Temporal Interest Points Chain (STIPC) for Activity Recognition", First Asian Conference on Pattern Recognition (ACPR), pp. 22-26, 2011.

[19] Mohamed Mansoor Roomi, Saranya, S.R. , Nashrin Banu, "Human action classification in partitioned feature space", IEEE International Conference on Machine Vision and Image Processing (MVIP), pp. 21-24, 2012.

[20] Jalal, A. , Uddin, M.Z. and Kim, T.-S. "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home", IEEE Transactions on Consumer Electronics, Volume: 58, pp. 863 – 871, 2012.

[21] Chunfeng Yuan1,XiLi, Weiming Hu1, Haibin Ling, "3D R transform on sptio temporal Interest points for action recognition" IEEE Conference on Computer Vision Pattern Recognition, pp. 724-730, 2013.