2020 IEEE International Conference for Innovation in Technology (INOCON)
Bangaluru, India. Nov 6-8, 2020

Self-Similarity Matrix and View Invariant Features Assisted Multi-View Human Action Recognition

Venkata Subbareddy K
Dept. of ECE
OUCE
Hyderabad, India
subvish@gmail.com

S. Sandhya Rani,
Dept. of CSE
MREC
Hyderabad, India
sarlanasandhya@gmail.com

***Abstract*--- This paper presents a novel Multi-View Human Action Recognition Framework which relies on Self Similarity Matrix (SSM), View and Rotation invariant features of human actions. In this paper, we accomplish the Self-Similarity between frames of an input video to extract only important frames. The self-similarity evaluation helps in the reduction of total number of training frames for every view. And further to extract scale and rotation invariant features, in this work we incorporate Gabor filter with varying scales and rotations. These features are learned through Support Vector Machine algorithm to recognize human actions under multiple views. Simulations are carried out over IXMAS multi-view dataset and the results are out-performing compared to the state-of-art methods.**

Keywords--- Human Action Recognition, Self-Similarity Matrix, Gabor Filter, IXMAS, Accuracy.

I. INTRODUCTION

Recently, Human Action Recognition (HAR) research is obtained a significant research interest due to its widespread applicability in various applications like Visual Surveillance, Video Retrieval, Entertainment, Human-Computer interaction, Autonomous vehicle driving etc. For example, a patient is undergoing a rehabilitation exercise at home, and his/her robot assistant is capable of recognizing the patient's actions, analyzing the correctness of the exercise, and preventing the patient from further injuries. Such an intelligent machine would be greatly beneficial as it saves the trips to visit the therapist, reduces the medical cost, and makes remote exercise into reality. The main objective of a HAR system is to identify the actions being performed in a video sequence under different situations such as occlusion, cluttering and different lighting conditions. The main center of this system is the computational algorithms which understand the human actions. Similar to the human vision system, these computational algorithms ought to produce a label after the analysis of partial or entire action in the video sequence [1], [2]. Developing such algorithms is typically addressed in the computer vision research, which studies how to make the computers to gain high level understanding regarding human actions from digital images and videos.

Recently, Multi-View Human Action recognition (MVHAR) methods have gained a significant interest due to the effective tackling capability of this system by the accomplishment of multiple cameras and observing an action in multiple views. MVHAR system is more robust than single view HAR system for view changes. MVHAR also considers the view point changes which has a significant impact on the action understanding. Hence the extraction of view invariant features from action video sequences is important. Based on this aspect, a novel MVHAR system is developed in this paper which is more effective in the extraction of view invariant features. The overall system architecture is shown in Figure.1. Initially, the key frames are selected for an action video based on self-similarity matrix (SSM). Further, towards the feature extraction, this paper accomplished Gabor filter in multiple orientations and also at various scales. After extracting the features, Support Vector Machine (SVM) algorithm is used for classification. Extensive simulations conducted over the developed system shows the outstanding performance with respect to the accurate action recognition for multiple views.

Rest of the paper is organized as follows; Section II illustrates the details of state-of-art techniques. Section III illustrates the details of proposed mechanism. Section IV illustrates the details of simulation experiments and finally the conclusions are provided in section V.

II. STATE-OF-ART

A. State of the art

Various solutions are developed in earlier for action recognition over years. Space-time shapes [3], covariance features [4], time evolution based human silhouettes [5], and local 3D patch descriptors [6] are some of the most popular techniques used for action representation. The further descriptors used for action representations are Space-time Interest Points (STIP) [7] [8] based, and Self-similarity matrices (SSM) [9] based approaches.

Recently, the SSM based action recognition has gained an effective action recognition performance. Imran Junejo et al., [9] introduced the concept of SSM. In this method, every action is represented with a set of SSMs. In this approach, initially, the action video is represented with low level features. Further the SSM is constructed based on the computation of Euclidean distance between the extracted low level features of all frames in a pairwise fashion.

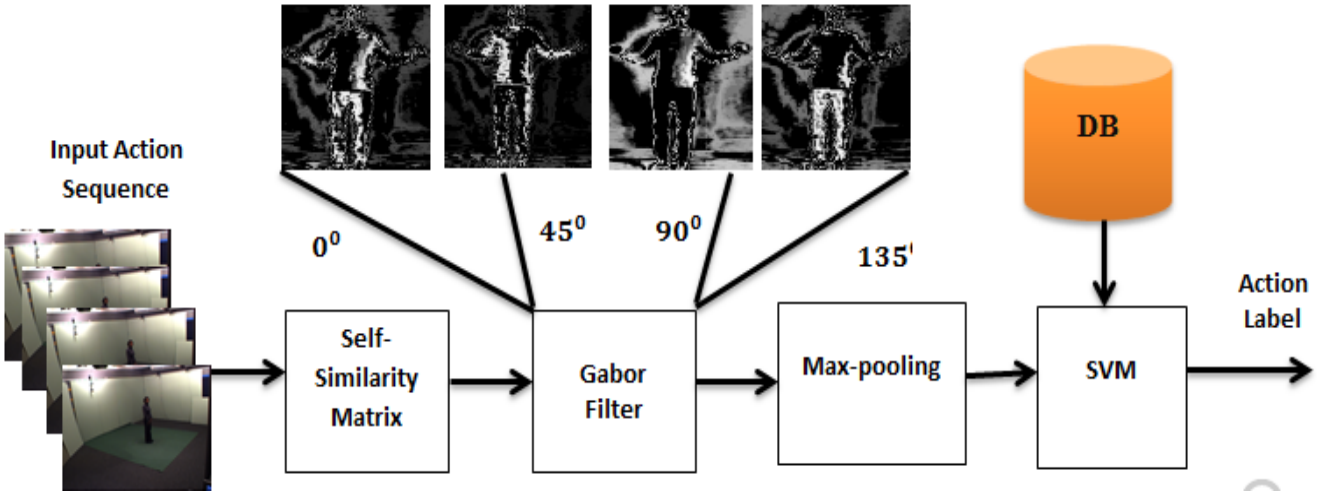


Fig. 1. Overall Schematic of proposed HAR system

Based on the SSM evaluation, a new variant of SSM called Local Self-Similarity (LSS) descriptor is introduced by E. Shechtman and M. Irani [10]), and Bag of Visual Words (BoVW) based LSS descriptor is introduced by Stark et al., [11]. However, the LSS never captures the global similarities in the entire image by which the image matching will be less effective.To overcome the LSS problem, T. Deselaers and V. Ferrari, [12] proposed the Global Self-Similarity (GSS) and explored its advantages over LSS. This captures the spatial arrangements of self-similarities within the entire image. This approach also had shown the effective utilization of GSS based descriptors to detect the objects through Branch and Bound Framework in a sliding window. Next, Jing Wang et al., [13] proposed a new HAR method based on SSM and Dynamic Time Warping (DTW). Here the SSM captures the Global Time information which was useful in the action recognition under viewpoints. Further, DTW is applied for the full pledged utilization of SSM information. K-Nearest Neighbor Classifier (KNNC) is accomplished for classifications.

In the feature extraction phase, considering the advantages of wavelet transform, A. Aryanfar et al., [14] proposed a novel method for MVHAR by integrating the wavelet transform [18] with silhouette. Initially, the contour of human silhouette is extracted and a distance signal is measured. In the next step, the wavelet transform is applied to extract the features of a single view and they are combined with features of multiple views. However, the wavelet transform is non-invariant to scaling due to the presence of down sampler. Next, Kuan Pen Chou et al., [15] proposed to extract the scale-invariant features and used to model the global spatial-temporal distribution. However this method is not robust for inter and intra class variations.

B. Problem Definition

In MVHAR system, Key frames selection is important to reduce the unnecessary computational time which incurs due to the processing of redundant frames at both training and testing phases. Most of the earlier approaches not focused in this direction. Furthermore, the extraction of both scale and rotation invariant features are also most important. Training only the rotation invariant features helps in the enhancement of recognition accuracy in multiple view-points.

III. KEY FRAMES SELECTION

In MVHAR, for every action, there exists multiple views and they are acquired through multiple cameras. However, in every view, only few frames are informative and remaining are redundant. This paper focused to select only those informative frames and tries to remove the redundant frames through the key frame selection process through SSM [9]. Given an Action video sequence F with N frames, $F = [F_1, F_2, \dots\dots\dots\dots\dots, F_N]$, a SSM is obtained as,

$$SSM(F) = \begin{bmatrix} 0 & e_{12} & e_{13} & e_{14} & \cdots & e_{1N} \\ e_{21} & 0 & e_{23} & e_{24} & \cdots & e_{2N} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ e_{N1} & e_{N2} & e_{N3} & d_{N4} & \dots & 0 \end{bmatrix} \quad (1)$$

Where $e_{ij} = \|p_i - p_j\|^2$ is the Euclidean distance between pixel intensities of frames F_i and F_j. Obviously the diagonal elements in the above matrix are zero which denotes a self-similarity between the same frames. Here the SSM is employed for all views and every time, ne view is considered as references. The simple pictorial representation of this process is shown in Figure.2

Over the obtained SSM Matrix, key frames are selected by finding the maximum differences (i.e., maximum e_{ij}). For example the first frame of first View is processed for subtraction from the first frame of remaining Views and among the obtained values, a minimum value is found. Based on that minimum values, last *N-1* frames are only selected which has lower minimum value, i.e., maximum difference ($\cong 0$). Simply, we select the frames which have maximum difference with first frame in the case of first frame as reference. Only one frame is excluded which have minimum difference with reference frame. This process is accomplished for second frame and also for further frames. Mathematically, it is performed as

$$[P, V] = \max(SSM(e_{i=1,\dots N, j=1,\dots N})) \quad (2)$$

Where P represents the position of frame which has minimum difference and V represents it's minimum value. In this manner, only few key frames are extracted from every View and they are only processed for feature extraction.

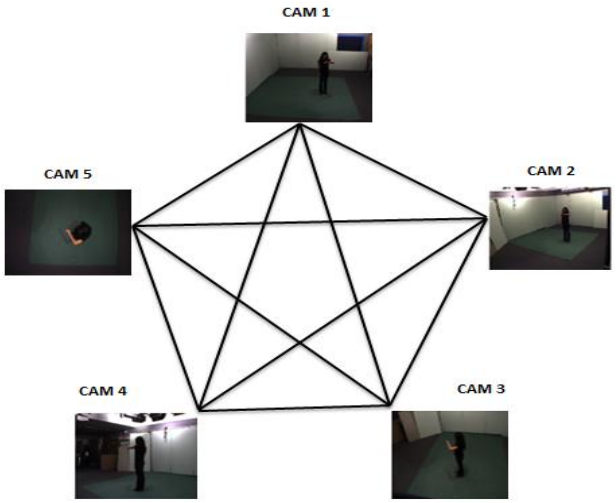


Fig. 2. Graphical representation of multiple views for SSM

IV. FEATURE EXTRACTION

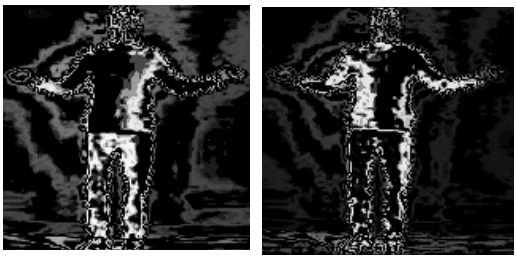
Towards features extraction, this work accomplished Gabor Filter due to its effectiveness in the feature extraction in different orientations. Since the Gabor filter extracts the features which are scale- and orientation-invariant, this paper considered it for Orientational features extraction. Here the Gabor filter is accomplished in four scales such as 5×5, 7×7, and 9×9, and four orientations such as 0^0, 45^0, 90^0, and 135^0. So totally for each frame, we will get $3 \times 4 = 12$ feature maps. The mathematical formula for Gabor filter is shown as

$$G(x,y) = exp\left(\frac{X^2+\gamma Y^2}{2\sigma^2}\right) \cos\left(\frac{2\pi}{\lambda}x\right) \quad (3)$$

Where

$$X = xcos\theta - ysin\theta, \;\; Y = xsin\theta + ycos\theta \quad (4)$$

Where (x, y) is position relative to the center of filter. The Gabor feature maps obtained at different orientations is shown in Figure.3. However the obtained 12 feature maps are high in number and create much computational burden.



(a) (b)

(c) (d)

Fig. 3. Gabor Feature maps at (a) 0^0 (b) 45^0 (c) 90^0 (d) 135^0

Hence, to reduce this computational burden, here a max pooling mechanism [16] is applied over the 12 feature maps of every frame. In other words, we will pick a maximum value from all feature maps with filter scale in each orientation. The max pooling in different scales is performed as

$$F_{max} = \max_{(x,y,\theta_s)}[F_{5\times5}(x,y,\theta_s), \dots, F_{9\times9}(x,y,\theta_s)] \quad (5)$$

Where F_{max} is the maximum feature map obtained through the max-pooling, $F_{k\times k}(x,y,\theta_s)$ is the feature map at $k \times k$ scale and at θ_s orientation. In this manner, we will get totally four feature maps, one from each orientation.

V. RESULTS AND DISCUSSION

For the simulation of proposed HAR system, comprehensive experiments are conducted over the well-known multi-view INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset [17]. IXMAS is a challenging dataset, acquired with multiple actors under multiple camera views. This dataset consists of 12 action classes such as Check Watch (CW), Cross Arms (CA), Scratch Head (SH), Sit Down (SD), Get Up (GU), Turn Around (TA), Walk (WK), Wave (WV), Punch (PH), Kick (KK), Point Out (PO) and Pick Up (PU). Each action is performed three times and 12 different subjects are recorded with five cameras, four are fixed at four sides and one is fixe on the top. These five cameras captures five views such as left, right front back and top. The frame rate is 23 frames per second and the size of frame is 390×291 pixels. Figure.4 shows some samples of different actions under multiple views. Each row represents different action and each column represents different views.

Table.1 shows the details of recognition performance at various cameras. From Table.1, we can observe that the obtained performance for Camera 4 is much better than the other cameras. The clear visibility of movements of all actions in Camera 4, the HAR system recognizes the actions more perfectly. Hence the training of actions through camera 4 is more important in the Multi-View Human Action recognition.

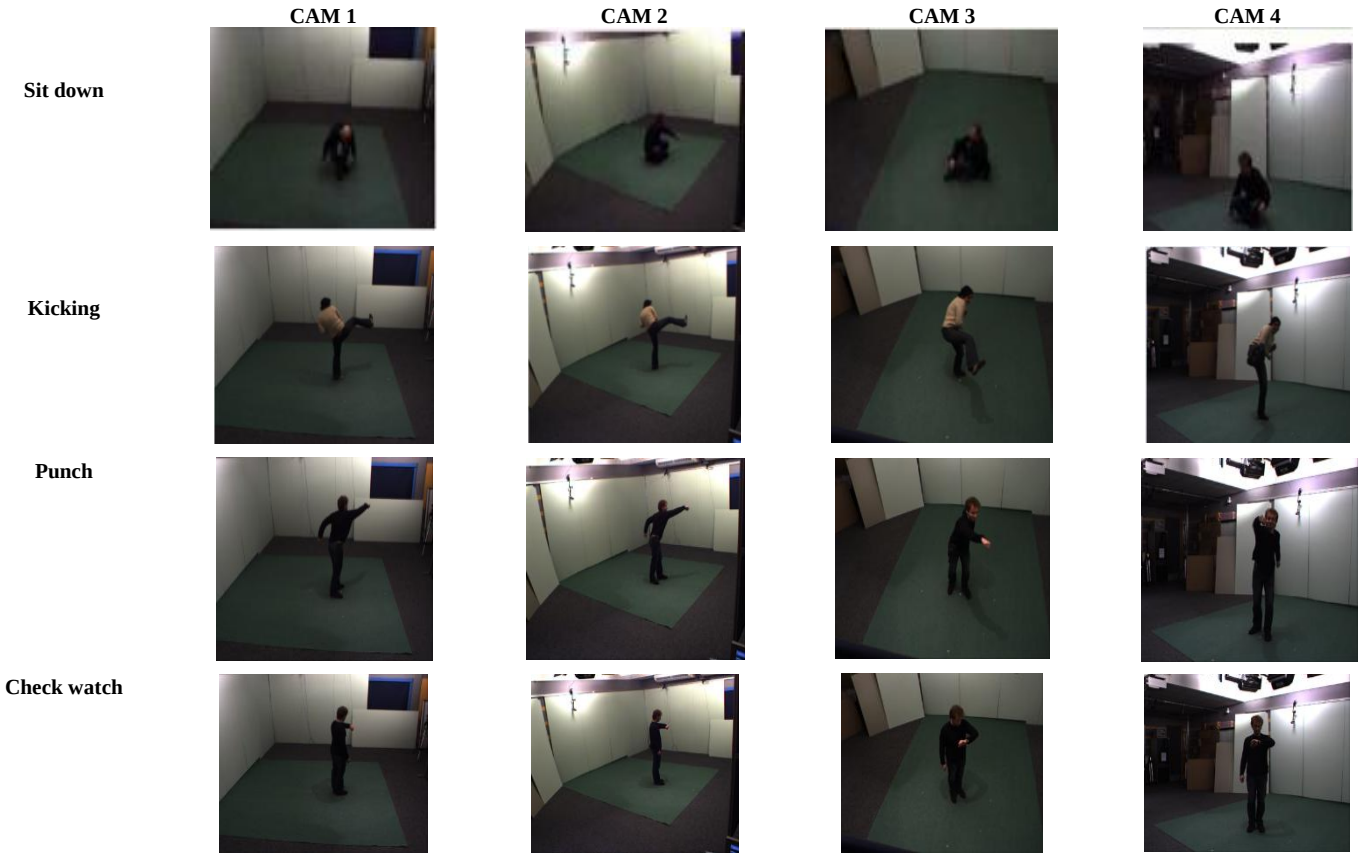


Fig. 4. Few samples of IXMAS dataset

TABLE I. CONFUSION MATRIX OF DIFFERENT CAMERAS

	Camera 1	Camera 2	Camera 3	Camera 4
Camera 1	**75.2356**	6.3321	8.5241	9.9082
Camera 2	14.2347	**66.3142**	10.2345	9.2166
Camera 3	9.6345	5.3363	**78.5523**	6.4769
Camera 4	9.8567	5.4278	4.5677	**80.1478**

TABLE II. CONFUSION MATRIX OF CAMERA 1

	CW	CA	SH	SD	GU	TA	WK	WV	PH	KK	PO	PU
CW	**85.1**	7.4	4.5	0	0	0	0	0	3.0	0	0	0
CA	20.2	**73.3**	3.25	0	0	0	0	3.25	0	0	0	0
SH	10.1	21.6	**64.1**	0	0	0	0	4.2	0	0	0	0
SD	0	0	0	**100**	0	0	0	0	0	0	0	0
GU	0	0	0	0	**100**	0	0	0	0	0	0	0
TA	0	0	0	0	0	**97.8**	2.2	0	0	0	0	0
WK	0	0	0	0	0	0	**100**	0	0	0	0	0
WV	6.8	10.2	6.8	0	0	0	0	**69.4**	6.80	0	0	0
PH	3.4	0	0	0	0	0	0	4.3	**92.3**	0	0	0
KK	0	0	0	0	0	0	0	0	1.2	**97.5**	1.3	0
PO	0	0	0	0	0	0	5.0	5.0	3.3	8.2	**78.5**	0
PU	0	0	0	0	0	0	0	0	0	0	0	**100**

TABLE III. CONFUSION MATRIX OF CAMERA 4

	CW	CA	SH	SD	GU	TA	WK	WV	PH	KK	PO	PU
CW	**80.3**	16.2	0	0	0	0	0	0	4.5	0	0	0
CA	18.1	**75.8**	0	0	0	0	0	0	4.1	0	2.0	0
SH	2.5	13.2	**72.1**	0	0	0	0	0	12.2	0	0	0
SD	0	0	0	**95.2**	0	0	0	0	4.8	0	0	0
GU	0	0	0	0	**85.4**	0	0	0	9.0	5.6	0	0
TA	0	0	0	0	0	**100**	0	0	0	0	0	0
WK	0	0	0	0	0	0	**100**	0	0	0	0	0
WV	6.4	6.3	21.1	0	0	0	0	**63.4**	4.4	0	0	0
PH	1.5	0	0	0	0	0	0	0	**86.4**	0	12.1	0
KK	0	0	0	0	0	0	0	0	0	**100**	0	0
PO	0	0	0	0	0	0	0	3.10	23.3	0	**73.6**	0
PU	0	0	0	0	0	0	0	0	0	0	0	**100**

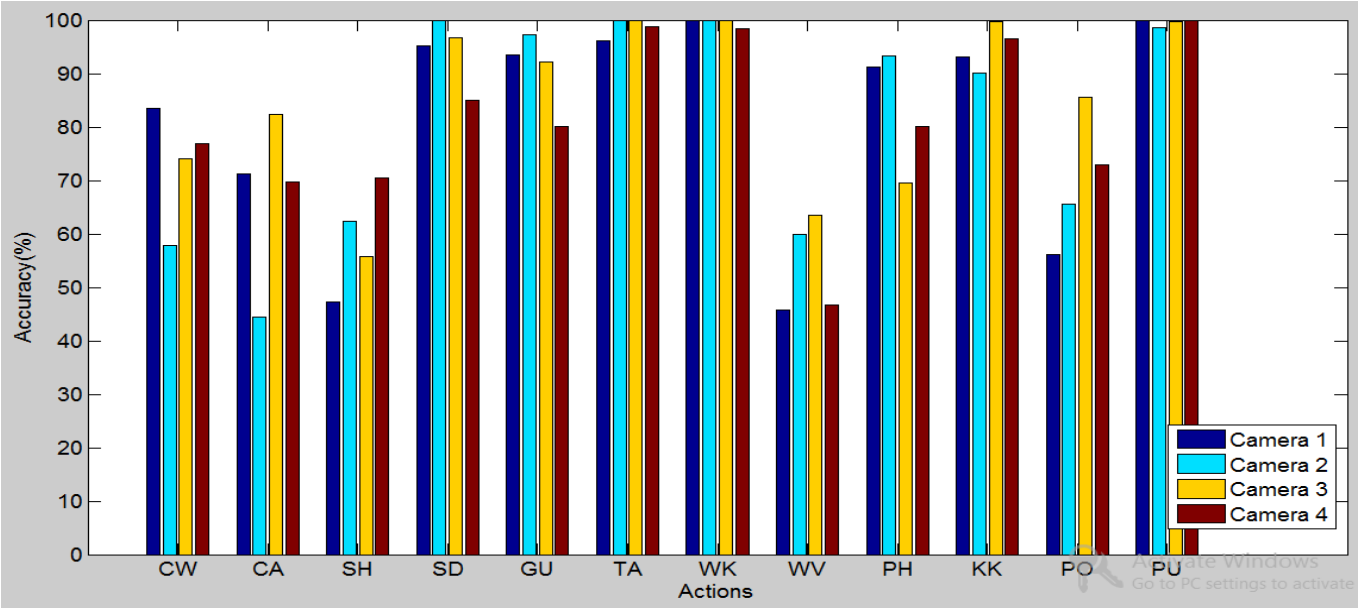


Fig. 5. Recognition performance of actions at different cameras

Further, Table.2 and Table.3 shows the confusion matrices of Camera 1 and Camera 4 respectively. From Table.2, we can notice that three actions such as Sit down, Get Up, Walk and Pick Up have achieved a maximum recognition performance (100%) and Scratch Head action has gained a minimum recognition rate (64.10%). In Camera 1, the actions are captured in the front view and this view explores the leg movement more clearly and the actions which are oriented to legs have maximum performance. Similarly from the Table.3, we can notice that totally four actions namely, Turn Around, Walk, Kick and Pick Up has gained a maximum recognition performance (100%) and the recognition performance is minimum is for wave action (63.40%). Due to the similar hand movements of Scratch Head and Wave, only 63.40% of actions are recognized as wave and 21.10% are recognized as Scratch Head. The similar observations can be observed for scratch head and point out actions.

The individual recognition accuracies of different actions is shown in Figure.5. We can observe that the proposed approach obtained a better performance for almost all actions. On an average, the actions namely sit down, get up, turn around, walk and Pick up has gained a maximum recognition performance at all cameras. The maximum recognition performance is obtained for Camera 4 and minimum is for Camera 1. In the actions captured through Camera 1, the majority of motions are blocked by human bodies and thus only slight movements are visible and hence for some actions like Scratch head and Wave captured under camera 1 have the less recognition rate. Moreover, we can also notice that the actions mainly having hand movements (e.g., cross Arms and check watch) have lower recognition rate than the actions those are associated with total human body parts movement (Turn-around and sit-down, pick up).

TABLE IV. COMPARISON THROUGH RECOGNITION ACCURACY (%)

Method/Camera	Camera 1	Camera 2	Camera 3	Camera 4	Average
A. Aryanfar et al. [14]	77.9189	79.5667	80.9220	83.2688	80.4191
Manish. Khare et al. [18]	77.4471	79.0949	80.4502	82.7970	79.9473
Proposed	80.6667	82.3145	83.6698	86.0166	**83.1669**

Table.4 shows the comparative analysis between the proposed and conventional approaches such as A. Aryanfar et al., [14] and M. Khare et al., [18]. We can observe that our performance is better than that of A. Aryanfar et al., [14] and M. Khare et al., [18]. In the conventional approaches, the Wavelet Transform is applied for feature extraction which has the main drawback of shift invariance. This problem is solved by Gabor filter which can provide rotation invariant features. Further to ensure the scale invariance, the action frames are processed for feature extraction under different scales. This process adds more effectiveness to the recognition performance. Moreover, the state-of-art methods not applied self-similarities for key frames extraction and the proposed have this is the only major advantage.

VI. CONCLUSION

In this paper, we focused over the problem of recognizing human actions under multiple views. For this purpose, we proposed a new MVHAR recognition framework based on SSM and Gabor filter. In the initial phase, this framework applied SSM over the action video to extract only key frames and then the obtained frames are processed for feature extraction through Gabor filter. Due to the rotation invariant features extraction from key frames, the action captured under multiple views is recognized effectively even with training of only one view. Simulation experiments are conducted over IXMAS MultiView dataset and observed promising results. On an average, the recognition accuracy of proposed approach is observed as 83.1669%. This approach achieved a significant improvement in the recognition accuracy in multiple view scenarios. In comparison to the state-of-art methods, our approach presents efficient results at all view instances.

REFERENCES

[1] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," IEEE Trans Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257–267, 2001.

[2] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos", in ICCV, 2011.

[3] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 984–989, 2005.

[4] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," IEEE Trans. Image Process., vol. 22, no. 6, pp. 2479–2494, 2013.

[5] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in Proc. 11th Eur. Conf. Comput. Vis., pp. 635–648, 2010.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–8, 2008.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 12, pp. 2247-2253, 2007.

[8] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape motion prototype trees," in Proc. 12th IEEE Int. Conf. Comput. Vis., pp. 444–451, 2009.

[9] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, no. PrePrints, 2010.

[10] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos." In CVPR, 2007.

[11] M. Stark, M. Goesele, and B. Schiele, "A shape-based object class model for knowledge transfer," In ICCV, 2009.

[12] T. Deselaers and V. Ferrari, "Global and efficient self-similarity for object classification and detection," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1633–1640, 2010.

[13] Jing Wang, and Huicheng Zheng, "View-robust action recognition based on temporal self-similarities and dynamic time warping," IEEE International Conference on Computer Science and Automation Engineering (CSAE), Zhangjiajie, China, 2012.

[14] A. Aryanfar, R. Yakob, and A. A. Halin, 2015, "Multi-View Human Action recognition Using Wavelet data Reduction and Multi-class Classification," In: Prof. of international Conf. on Soft Computing and Software Engineering, Berkeley, Suta, pp.585-592, 2015.

[15] K. P. Chou, M. Prasad, D. Wu , N. Sharma, D. L. Li, Y. F. Lin, M. Blumenstein, W. C. Lin, And C. T Lin, "Robust Feature-Based Automated Multi-View Human Action Recognition System," IEEE Access, Vol. 6, pp.15283-15296, 2018.

[16] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex", Natural Neuroscience, Vol.2, pp. 1019–1025, 1999.

[17] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," Computer Vision and Image Understanding, vol. 104, no. 2-3, 2006.

[18] Manish Khare and Moongu Jeon, "Towards Discrete Wavelet Transform-based human activity recognition", Proc. SPIE 10443, Second International Workshop on Pattern Recognition, Vol. 19, 2017.