Murtaza, F., Yousaf, M.H. y Velastin, S.A. (2015). Multi-view Human Action Recognition Using Histograms of Oriented Gradients (HOG) Description of Motion History Images (MHIs). *In 2015 13th International Conference on Frontiers of Information Technology (FIT),* pp. 297-302.

DOI: https://doi.org/10.1109/FIT.2015.59

Multi-view Human Action Recognition using Histograms of Oriented Gradients (HOG) Description of Motion History Images (MHIs)

Fiza Murtaza
Department of Computer Engg.
University of Eng. & Tech.
Taxila, Pakistan
fizamurtaza@yahoo.com

Muhammad Haroon Yousaf
Department of Computer Engg.
University of Eng. & Tech.
Taxila, Pakistan
haroon.yousaf@uettaxila.edu.pk

Sergio A. Velastin
Department of Computer Science
University Carlos III de Madrid
Spain,
sergio.velastin@ieee.org

***Abstract*—In this paper, a silhouette-based view-independent human action recognition scheme is proposed for multi-camera dataset. To overcome the high-dimensionality issue, incurred due to multi-camera data, the low-dimensional representation based on Motion History Image (MHI) was extracted. A single MHI is computed for each view/action video. For efficient description of MHIs Histograms of Oriented Gradients (HOG) are employed. Finally the classification of HOG based description of MHIs is based on Nearest Neighbor (NN) classifier. The proposed method does not employ feature fusion for multi-view data and therefore this method does not require a fixed number of cameras setup during training and testing stages. The proposed method is suitable for multi-view as well as single view dataset as no feature fusion is used. Experimentation results on multi-view MuHAVi-14 and MuHAVi-8 datasets give high accuracy rates of 92.65% and 99.26% respectively using Leave-One-Sequence-Out (LOSO) cross validation technique as compared to similar state-of-the-art approaches. The proposed method is computationally efficient and hence suitable for real-time action recognition systems.**

Keywords—action recognition; histograms of oriented gradients (HOG); motion history images (MHIs); MuHAVi dataset

I. INTRODUCTION

For understanding visual environment it is a critical task to develop effective and automatic methods to analyze video data efficiently. There is a great demand for computer vision methods to process, analyze and understand videos in an automatic mode which can be possible by human activity recognition. Human activity Recognition is one of the promising domain in computer vision due to its multi-dimensional applications in Human-Computer-Interaction, sports monitoring, content based video search and retrieval, video surveillance, wild life monitoring etc. [1]. Depending upon the duration and complexity, the human activities are categorized into gestures, primitive actions, interactions and group activities [2]. Gestures are the simple movement of human body parts, while the primitive actions are made up of multiple gestures. Interactions are the actions of multiple subjects/objects and group activities are the activities performed by group of people.

Single-view human action recognition has been well studied in the last few decades. In single-view human action recognition there are approximately three types of features used for action representation used in literature. These are holistic, local features, and geometric human body features.

Holistic approaches use shape [3] or motion based [4] information. Shape based methods do not depend upon a moving person's clothes color, texture, and luminous therefore these are desirable for better action representation. Motion based approaches are unpredictable in case of motion discontinuities, low quality videos and small variations in background. In geometric human body features, action recognition is done on the basis of body part locating along with movements [5]. Local space-time features or interest points [6] compute the shape and motion characteristics in video locally and feature descriptor is used [7] to describe these features efficiently. These features are robust against scale, background motion and clutter. These features do not require object segmentation, hence they are extracted directly from video frames.

As single-view approaches uses one camera for capturing the human action, therefore these approaches supposed to have same camera-view during training and testing. If this condition is not met, the accuracy of these approaches significantly decreases because same actions look quite different when captured by arbitrary viewing angles [8]. Therefore view-invariant action recognition is presented in the last decade by exploiting multi-camera for better action description, as surveyed in [9], [10].

Multi-view approaches are divided in two categories [11]; 3D and 2D multi-view methods. 3D multi-view methods [12]-[15] represent the actions by using 3D fused data therefore, these methods, require a (fixed) multi-camera setup during training as well as testing stages. This is a limiting application setup because in real-time scenarios the actor under attention is not required to be visible in all cameras either because actor is outside the range of camera setup or due to occlusion [16].

To overcome this limitation, different types of directions using 2D multi-view methods have been proposed by researchers. The first direction is based on a single-view view-independent approach. In this approach, action recognition is accomplished on every video independently coming from all the cameras of network. View-invariant action representation is proposed in [17]-[19] and classification is carried out by training a universal classifier for all available views or by using multiple classifiers for training [20]. Final results are obtained on the basis of fusion of results from all classifiers. The second direction is based on multi-view learning during training and testing of unknown action is done based on these learned features. As in [21]-[23] no feature fusion is used therefore there is no

need to have all camera views available during training stage. The advantage of their approach is that they can handle missing views of an action. The third direction is based on cross-view action recognition. In this approach one view is used for training while other for testing. Many techniques have been anticipated using cross-view action recognition such as transfer learning [24], [25], information maximization [26] etc.

Using a multi-view dataset, view-independent action recognition can be achieved at the cost of higher processing time and storage requirements. However higher processing time is not affordable in real time action recognition systems. In this paper, a silhouette-based view-independent human action recognition method is proposed. To overcome the high dimensionality issue incurred due to multi-camera action representations, this paper is focuses on the low-dimensional representation of multi-view data based on 2D motion template. For action representation Motion History Images (MHIs) [27] for each view/action video is computed, which encodes how recently motion occurred at a pixel. For efficient description of MHIs Histograms of Oriented Gradients (HOG) [28] is employed. A Nearest Neighbor (NN) classifier is used for action classification. The proposed method outperforms similar state-of-art approach on well-known benchmark multi-view dataset, MuHAVi [29]. This work focuses on the recognition of low-level gestures and actions; therefore this task is called action recognition.

The rest of the paper is organized as follows: Section II presents an overview of proposed multi-view human action recognition approach. Section III presents description of experimental results on benchmark datasets. Finally, conclusions are provided in Section IV.

II. PROPOSED METHODOLOGY

The 2D motion based templates, based on MHIs as proposed by [27], are used to capture and model human actions. Moreover, it is assumed that the silhouette images are available. The multi-camera human action recognition framework is presented in Fig. 1.

In training mode MHIs are calculated from the silhouettes for every action/views. Then these MHIs are centered with respect to the center of mass of detected object in order to have location independent representation. Then the resulted MHIs are scaled to some fixed size in order to have scale invariant representation. After calculating MHIs, the HOG descriptor is computed over all available MHIs. Then the NN classifier is trained over HOG description of the MHIs. In testing mode the same steps are repeated for input query video. For the classification, an NN classifier is used to find the best match of the query video and the corresponding class label of the best matched sequence will be assigned to the query video.

In the next sections, the computation of MHI, its efficient description using HOG descriptor and NN classifier is described in more detail.

A. *MHI template*

The observation behind motion templates is that a human action makes a space-time shape in a space-time volume in an action video. One action sequence can be described by a single MHI image. MHI arrests information of motion by encoding how motion has changed at a pixel. Let say I(x, y, t) is a binary image and if I(x, y, t) = 1 then this shows that there exist motion (foreground) at time t at location (x, y). The MHI at time t is computed as:

$$MHI_t = \begin{cases} \tau & \text{if } I(x,y,t)=1 \\ \max\left(0, MHI_{t-1}(x,y) - 1\right) & \text{otherwise} \end{cases} \quad (1)$$

where τ is the number of frames used for computing the MHI template. In all experiments, the value of τ is taken to be equal to total number of frames in an action video. The resulting MHI can represent the motion sequence in a compact manner. The silhouette sequence belonging to one action is compressed into a gray scale image, where most recent foreground is represented by brighter gray-value pixels. It preserves dominant motion information.

MHIs are centered with respect to the center of mass of the detected foreground and scaled to some fixed size in order to have a scale and location invariant representation, as shown in Fig. 2. An illumination and contrast invariant representation MHI_n can be achieved by normalization:

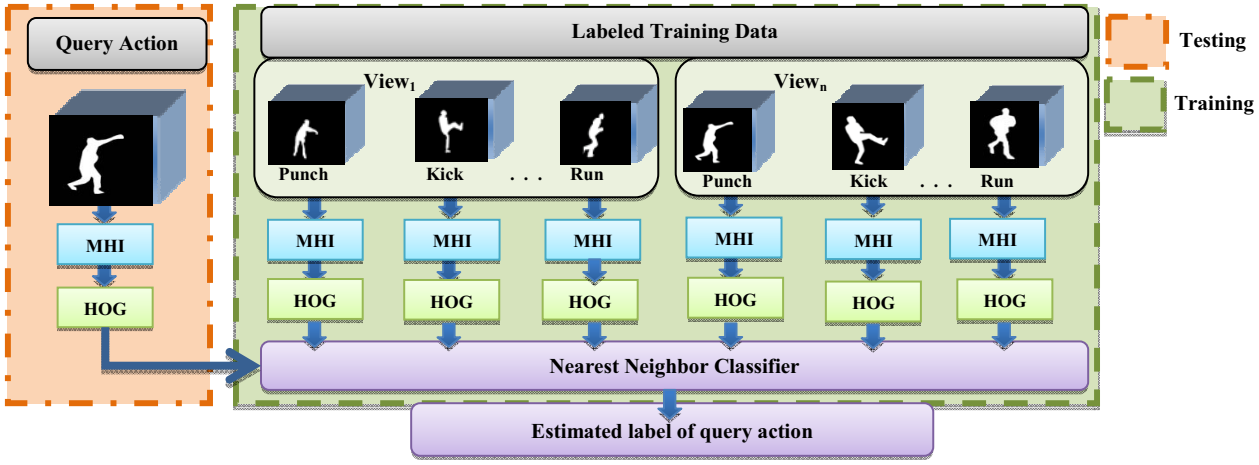

Figure 1. General overview of multi-view human action recognition framework

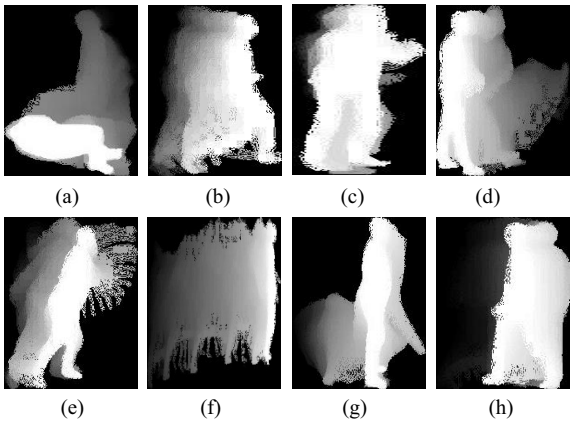

Figure 2. Examples of MHIs from MuHAVi-14 dataset of action classes (a) CollapseRight, (b) GuardToKick, (c) GuardToPunch, (d) KickRight, (e) PunchRight, (f) RunLeftToRight, (g) StandupLeft (h) TurnBackLeft

$$MHI_n = \frac{MHI_\tau}{\sum_{x=1}^{M}\sum_{y=1}^{N} MHI_\tau(x,y)} \quad (2)$$

where M is the total number of rows, N is the total number of columns in MHI_τ.

View-point dependencies can be removed by generating a 3D model of human body as in [12]. However, they used a fixed number of many calibrated cameras during training as well as testing phase, which is not possible in real time scenarios. To overcome this limitation, our method uses multiple cameras during training to obtain MHIs for each view independently. Thus we do not need calibrated cameras and feature fusion of multiple views to handle missing camera views. In real time applications it is not necessary that a subject is being observed from all cameras, therefore, the proposed approach is suitable for such scenarios.

The MHI representation is less sensitive to noise, shadows, missing body parts and holes in the foreground objective arising from imperfect segmentation. The MHI is sensitive to direction of motion of action; therefore, it can discriminate between actions of opposite directions (e.g. run left and run right). The MHI can encode a range of times in a final and single, frame as used in this work, which make this method more computationally efficient.

B. HOG based description of MHIs

HOG feature descriptors [28] are widely used in computer vision for object detection and human action recognition. MHIs can be regarded as "saliency masks" on which a more robust descriptor such as HOG is used (in contrast to those approaches that use MHIs on their own such as [21]). Here, the HOG descriptor is computed over MHIs (as shown in Fig. 3) to obtain discriminant description of MHIs for better action recognition:

1) Gradient computation: First the horizontal and vertical gradients are obtained by filtering the MHI image by the following kernels in x and y direction:

$$D_x = [-1\ 0\ 1]\ and\ D_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}. \quad (3)$$

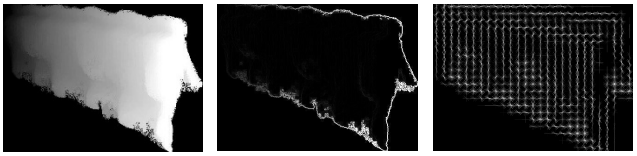

Figure 3. HOG implementation (a) MHI of action "WalkLeftToRight" (b) gradient magnitude and (c) orientation binning visualization

After horizontal and vertical gradients (I_x ,I_y) calculation the magnitude and orientation of the gradient magnitude G and orientation θ are calculated by:

$$|G| = \sqrt{{I_x}^2 + {I_y}^2}\ . \quad (4)$$

$$\theta = \tan^{-1}\left(\frac{I_x}{I_y}\right)\ . \quad (5)$$

The resultant orientation values, from (5), are between [-180°, 180°]. Unsigned orientations increase the performance therefore it is desirous for this implementation, the angles less than 0° are summed up with 180° the have unsigned values of orientation.

2) Orientation binning: The histograms of cells are computed which will be used for descriptor blocks. Non-overlapping cells of size 8x8 pixels are computed over gradient image. 9 orientation bins are used for histogram computation on the interval of [0°, 180°]. For each pixel, its corresponding bin is found on the basis of its orientation and its related magnitude, as in (4) and (5), is voted into this bin.

3) Descriptor blocks: In order to have illumination and contrast invariance representation the cells must be grouped into larger and connected blocks. There are two block geometries, rectangular R-HOG and the circular C-HOG. In the implementation R-HOG geometry is used in which each R-HOG block has 2x2 cells. Final descriptor is formed by concatenation histograms of all cells into a row or column vector.

C. NN classifier

The proposed method employed simple yet effective NN classifier approach to obtain nearest class label. Amongst the various approaches to supervised learning classifiers, the NN classifier achieves reliably high accuracy rates. NN classifier does not require any *priori* assumptions about the distributions of the training examples of data. The idea of NN classifier is easy and direct: test sample is given to classifier, the classifier will find the most similar sample (under some distance measure) in the labeled training data and finally the label of most similar sample is assigned to test sample. We have used an Euclidian metric for distance calculation using:

$$d = \sqrt{\sum_{j=1}^{n} (p_j - q_j)^2} \quad (6)$$

where p and q are the HOG vectors and n is the length of vector. The test sample is compared with each training sample and the label of the sample having minimum distance is assigned to test sample.

III. EXPERIMENTATION RESULTS AND DISCUSSION

The efficiency of algorithm is tested on MuHAVi dataset [29]. Some of the examples of our proposed method are shown in Fig. 4, in which video from the specific action is changed into its specific MHI representation and then its HOG description is obtained. Next, the MuHAVi dataset and validation schemes used for the proposed algorithm are explained.

A. MuHAVi Dataset

MuHAVi consists of multiple cameras human action data. It contains 17 actions performed three to four times by 14 persons. 8 CCTV cameras are used to capture execution of actions; these cameras are mounted at 45^{o} apart at the four corners and centers of rectangular action region. This dataset also contains manually annotated silhouette set known as MuHAVi-MAS, where there are 136 manually annotated sequences containing 14 primitive actions performed by two actors captured by two viewing angles. 14 primitive actions are also called as MuHAVi-14 and theses primitive actions can be merged into 8 action classes. In MuHAVi-8; e.g. "RunLeftToRight" and "RunRightToLeft" can merged into one action class "Run". Different experiments based on different validation schemes are performed on MuHAVi for performance evaluation.

Three types of cross validation schemes; leave-one-sequence-out (LOSO), leave-one-actor-out (LOAO), leave-one-camera-out (LOCO) cross validation are used on MuHAVi. For all of the experiments in this section, the value of τ is set to be equal to total number of frames in an action, 9 orientation bins and 8x8 cell sizes for HOG descriptor is used. Experimental results and comparisons, using different validation schemes, are discussed in the following sections;

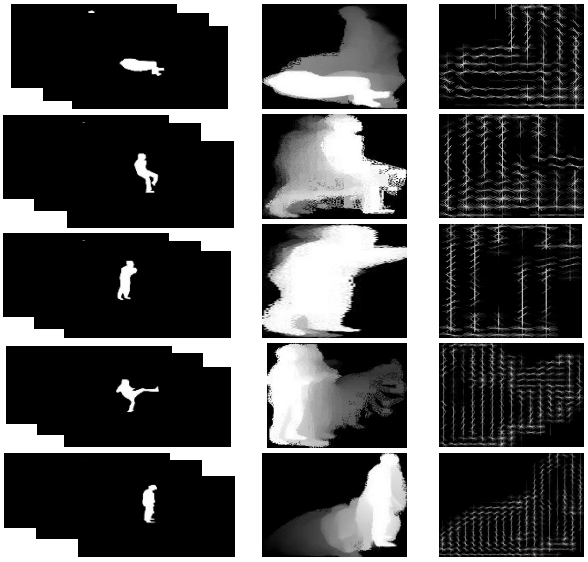

Figure 4. Demonstration of our proposed method on few actions from MuHAVi. Column1: "Collapse", "GuardToKick", "GuardToPunch", "Kick" and "StandUp" actions and their corresponding MHI and HOG descriptions are shown in Column2 and 3 respectively.

1) LOSO cross validation: In LOSO the KNN classifier is trained on all sequences except one and that left-one is used for testing (135 and 1 respectively). This is repeated for all possible combinations to compute averages.

The proposed method attains promising accuracy rates of 99.26% for MuHAVi-8 and the resultant confusion matrix is shown in Fig. 5. Only one sequence is misclassified i.e. "TurnBack" action is confused with "Walk" action. Thus reveals that the proposed algorithm is more robust and efficient in circumstances where an action is made up of multiple primitive actions e.g. "CollapseLeft" and "CollapseRight" combined into "Collapse".

The proposed method attains high accuracy rates of 92.65% for MuHAVi-14 and the resultant confusion matrix is shown in Fig. 6. Notice that "GaurdToKick" action is misclassified with "GaurdToPunch" due the similarity between these two actions. Other than "StandUpLeft/Right" action, it is worth notice (from Fig. 6) that MHI is sensitive to direction of motion of action; therefore, it can easily discriminate between actions of opposite directions e.g. "WalkLeftToRight" and "WalkRightToLeft".

Our proposed method outperforms the similar state-of-the-art approaches [22] and [29] except [23] (recorded in TABLE I), but [23] did not evaluated their algorithm on MuHAVi-8 also they have used only one cross validation scheme i.e. LOSO.

2) LOAO cross validation: MuHAVi dataset contains two actors, in LOAO cross validation scheme the experiment is performed by training the classifier on sequences of one actor testing on the sequences of second actor. The average accuracy is calculated by alternatively testing both actors.

The results (TABLE I) reveal that the proposed approach is more robust against the variability of actors' style of doing an activity as the training is performed on one actor and testing on another one and vice versa. This experiment got accuracy rates of 81.62% and 94.12% for MuHAVi-14 and MuHAVi-8 respectively. Our proposed

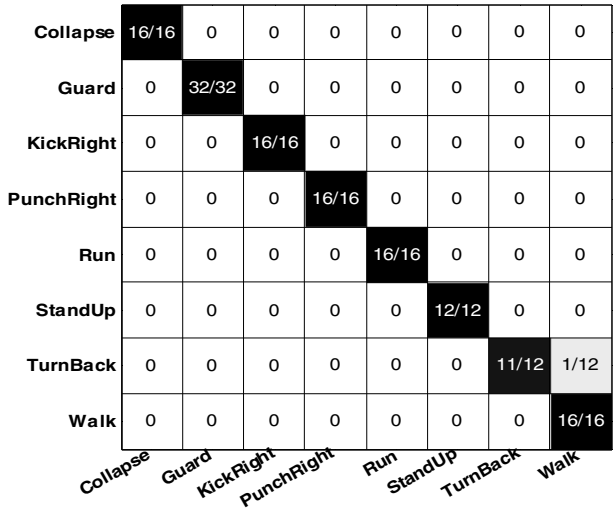

Figure 5. Confusion matrix for MuHAVi-8 based on LOSO (average accuracy = 99.26%)

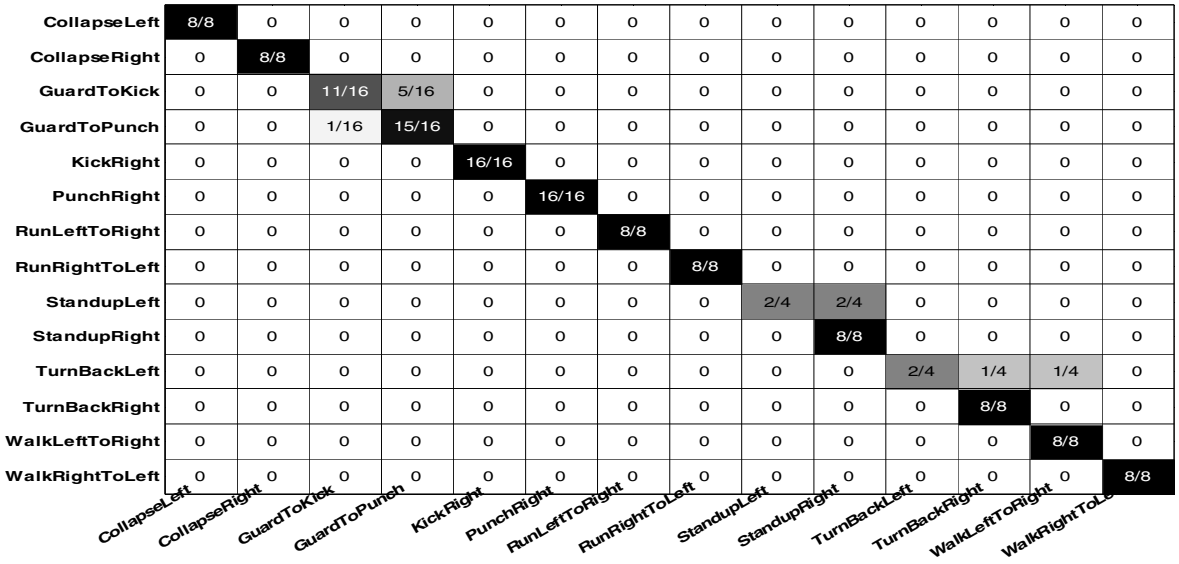

Figure 6. Confusion matrix of MuHAVi-14 based on LOSO (average accuracy = 92.65%)

method outperforms the similar state-of-the-art approaches [21], [22] and [29] (recorded in TABLE I). Using LOAO cross validation scheme, as far as we know, these are the highest accuracy rates achieved so far on MuHAVi-8 dataset.

3) LOCO cross validation: In LOCO cross validation schemes the classifier is trained on all sequences taken from one camera and tested on another set of sequences belonging to another camera. MuHAVi contains two camera-views available therefore average accuracy is calculated by testing alternatively on both of the available camera-views. This validation scheme presents robustness of the proposed algorithm towards change in viewpoint.

This experiment got accuracy rates of 54.41% and 75.74% for MuHAVi-14 and MuHAVi-8 respectively. It is clear from this experiments that the proposed approach is robust against the viewpoint changes as the training is performed on one camera and testing on another one and vice versa. As MuHAVi consists of only two camera views, accuracy can be increase using more training camera-views.

Again our proposed method outperforms the similar state-of-the-art approaches [22] and [29] (recorded in TABLE I). Using LOCO cross validation scheme, as far as we know, these are the highest accuracy rates achieved so far on MuHAVi-14 and MuHAVi-8 datasets.

B. Temporal evaluation

All experiments are performed on MATLAB 7 with Intel core i3 at 1.70 GHz, 4GB RAM, 64 bit operating system. In MuHAVi, there are 136 sequences composed of 7941 frames having resolution of 720 x 576 pixels. For MuHAVi-14, training and testing of 136 sequences takes 47.18 s, i.e. on average it takes speed of 0.35 s per sequence at the rate of 168.3 Frames/Second (FPS). While for MuHAVi-8, the training and testing process for the whole dataset (136 sequences) takes 46.8 s, at an average speed of 0.34 s per sequence at the rate of 169.7 FPS.

More FPS rate for MuHAVi-8 compared to MuHAVi-14 is due to the fact that it has less number of classes than MuHAVi-14. For MuHAVi-8/14 [22] achieved 56FPS, which shows that our proposed approach is about 3 times faster than their method. Whereas [21] achieved only 3.4FPS which limit their method for real time action recognition of human actions. Our proposed method is about 50 times faster than [21].

TABLE I. Comparison results on the basis of recognition accuracy

Approach	Accuracy (%)		Cross Validation Scheme
	MuHAVi-14	MuHAVi-8	
Cheema et al. [22]	86.03	95.58	LOSO
Cai et al. [23]	**98.53**	N/A	
Singh et al. [29]	82.35	97.80	
This paper	**92.65**	**99.26**	
Orrite et al. [21]	75.00	85.94	LOAO
Cheema et al. [22]	75.53	83.08	
Singh et al. [29]	61.76	76.47	
This paper	**81.62**	**94.12**	
Cheema et al. [22]	50.0	57.4	LOCO
Singh et al. [29]	42.6	50.0	
This paper	**54.41**	**75.74**	

These tests include all the steps of computing MHIs, HOG description and the final recognition process using KNN classifier, and assuming that silhouette images are available for all experiments. TABLE II shows the final comparison of proposed approach with the state-of-the-art approach [21] and [22] using LOAO cross validation scheme. In this paper, a silhouette-based view-independent human action recognition method is proposed for multi-camera dataset. HOG based description of MHIs has resulted in high accuracy rates using simple NN classifier as compared to similar state-of-art methods. The NN classifier is adopted for simplicity. Probably the use of more

sophisticated classifiers would improve these results, but the relatively high accuracy of the simpler NN classifier approach supports the appropriateness of choosing HOG descriptor.

TABLE II. Comparison results based on recognition accuracy and speed

Approach	MuHAVi-14		MuHAVi-8	
	Accuracy %	FPS	Accuracy %	FPS
Orrite et al. [21]	75.00	2.1	85.94	3.4
Cheema et al. [22]	75.53	56	83.08	56
This paper	**81.62**	**168.3**	**94.12**	**169.7**

IV. CONCLUSION

The proposed method can cope with high-dimensionality data incurred, which is achieved by using a single MHI for the image sequence of each action. As no feature fusion is applied, the proposed method is suitable for real-time scenarios because the subject is not necessarily available from all camera-views. The proposed method does not employ feature fusion for multi-view data therefore this method does not require a fixed number of cameras during training and testing. The proposed method achieved high accuracy rates and is computationally efficient as compared to state of the art techniques. The high processing speed of the proposed method makes it suitable for real-time action recognition systems.

For the future work, clustering techniques can be used to obtain key MHI poses, which may describe an action with a more distinctive representation. Along with clustering techniques, feature descriptors other than HOG descriptor can be used.

ACKNOWLEDGEMENTS

S.A. Velastin acknowledges funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement nº 600371, el Ministerio de Economía y Competitividad (COFUND2013-51509) and Banco Santander.

REFERENCES

[1] P. Ronald. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976-990, 2010.

[2] J. K. Aggarwal, and Michael S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)* 43.3 (2011): 16.

[3] M. Ahmad, I. Parvin, and S. W. Lee. (2010). Silhouette history and energy image information for human movement recognition. *Journal of Multimedia*, 5(1):12-21.

[4] H. Wang, A. Klaser, C. Schmid, and C. Liu. (2011). Action recognition by dense trajectories. In Proceedings of the IEEE International Conference on *Computer Vision and Pattern Recognition*. pp. 3169-3137.

[5] D. Ramanan, D.A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. IEEE T-PAMI, 29:65-81, 2007.

[6] Wong, S. F. and Cipolla, R. (2007). Extracting spatio-temporal interest points using global information. In Proceedings of the IEEE *International Conference on Computer Vision*, pp. 1-8.

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91-110, 2004.

[8] D. Rudoy and L. Z. Manor. Viewpoint Selection for Human Actions. In *IJCV*, 2012. pp. 243-254.

[9] X. Ji and H. Liu. Advances in view-invariant human motion analysis: A review. Trans. *Sys. Man Cyber*, 2010. pp. 13-24.

[10] M. B. Holte, T. B. Moeslund, C. Tran and M. M. Trivedy. Human Action Recognition using Multiple Views: A Comparative Perspective on Recent Developments. *ACM HGBU*, 2011. pp. 47-52.

[11] A. Iosifidis, A. Tefas and I. Pitas, Multi-view human action recognition: A survey. *Intelligent Information Hiding and Multimedia Signal Processing,* 2013. pp. 522-525.

[12] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2–3):249–257, 2006.

[13] M. Holte, T. Moeslund, N. Nikolaidis, and I. Pitas. 3D human action recognition for multi-view camera systems. *3DIMPVT*, 2011, pp. 342-349.

[14] P. Yan, S. Khan, and M. Shah. Learning 4D action feature models for arbitrary view action recognition. *CVPR*, 2008. 1-7

[15] N. Gkalelis, N. Nikolaidis, and I. Pitas. View independent human movement recognition from multi-view video exploiting a circular invariant posture representation. *IEEE ICME*, 2009. pp. 394-397.

[16] F. Qureshi and D. Terzopoulos. Surveillance camera scheduling: A virtual vision approach. *Multimedia Systems*, vol. 12, no. 3, pp. 269-283, 2006. pp. 269-283.

[17] F. Zhu, L. Shao and M. Lin. Multi-view action recognition using local similarity random forests and sensor fusion. *Pat. Rec. Letters*, vol. 24, pp. 20–24, 2013, pp. 20-24.

[18] A. Iosifidis, A. Tefas and I. Pitas. View-Invariant Action Recognition Based on Artificial Neural Networks. *IEEE TNNLS*, vol. 23, no. 3, pp. 412–424, 2012.

[19] A. Iosifidis, A. Tefas and I. Pitas. Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis. *Sig. Proc*., vol. 93, no. 6, pp. 1445–1457, 2013.

[20] M. Ahmad and S. W. Lee. HMM-based human action recognition using multiview image sequences. *ICPR*, 2006, pp. 263-266.

[21] C. Orrite, M. Rodriguez, E. Herrero, G. Rogez, S.A. Velastin. Automatic Segmentation and Recognition of Human Actions in Monocular Sequences. *ICPR,* 2014, pp. 4218-4223

[22] S. Cheema, A. Eweiwi, C. Thurau, C. Bauckhage. Action recognition by learning discriminative key poses. *ICCV Workshops,* 2011, pp. 1302-1309.

[23] J. Cai, X. Tang, G. Feng. Learning Pose Dictionary for Human Action Recognition. *ICPR*, vol., no., pp.381-386, 2014.

[24] R. Li. Discriminative virtual views for cross-view action recognition. *CVPR*, 2012, pp. 2855-2862.

[25] B. Li, O. I. Campls and M. Sznaier. Cross-view Activity Recognition using Hankelets. *CVPR*, 2012, pp. 1362-1369

[26] J. Liu and M. Shah. Learning human actions via information maximization. *CVPR*, 2008, pp. 1-8.

[27] A. Bobick and J. Davis. The recognition of human movement using temporal templates. IEEE Trans. *Pattern recognition and Machine Intelligence*, 23(1):257– 267, 2001.

[28] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886-893, 2005.

[29] S. Singh, S.A. Velastin and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *Proc. AVSS*, 2010., pp. 48-55.