Fiza Murtaza, Muhammad Haroon Yousaf, Sergio A. Velastin. (2016). Multi-view Human Action Recognition using 2D Motion Templates based on MHIs and their HOG Description. *IET Computer Vision*, 10 (7), pp. 758-767, Oct. 2016.

DOI:10.1049/iet-cvi.2015.0416

Multi-view Human Action Recognition using 2D Motion Templates based on MHIs and their HOG Description

Fiza Murtaza[1], Muhammad Haroon Yousaf [1], Sergio A. Velastin [2,3] *

[1] Department of Computer Engg., University of Engg. & Tech. Taxila, Pakistan
[2] Department of Computer Science and Engineering, University Carlos III de Madrid, Spain
[3] Department of Informatic Eng., University of Santiago, Santiago, Chile
* Corresponding author (sergio.velastin@theiet.org)

Abstract: In this paper, a new multi-view human action recognition approach is proposed by exploiting low dimensional motion information of actions. Before feature extraction, pre-processing steps are performed to remove noise from silhouettes, incurred due to imperfect but realistic segmentation. 2D motion templates based on Motion History Image (MHI) are computed for each view/action video which can cope with the high-dimensionality issue, incurred due to multi-camera data. Histograms of Oriented Gradients (HOGs) are used as an efficient description of the MHIs. Finally, a Nearest Neighbor (NN) classifier is employed for the classification of the HOG based description of MHIs. As compared to existing approaches, the proposed method has three advantages: 1) does not require a fixed number of cameras setup during training and testing stages hence missing camera-views can be tolerated, 2) requires less memory and bandwidth requirements and hence 3) is computationally efficient which makes it suitable for real-time action recognition. The proposed method is evaluated on the new MuHAVi-uncut dataset having a large number of action categories and a large set of camera-views with noisy silhouettes. As far as we know, this is the first report of results on that dataset and can be used by future workers as a baseline to improve upon. Experimentation results on multi-view with this dataset gives a high accuracy rate of 95.4% using Leave-One-Sequence-Out (LOSO) cross validation technique and compares well to similar state-of-the-art approaches.

1. Introduction

For understanding visual environments 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 for human activity recognition (HAR). HAR is one of the promising domains in computer vision due to its many applications in Human-Computer-Interaction, sports monitoring, content based video search and retrieval, video surveillance, wild life monitoring etc. [1]. Depending upon duration and complexity, human activities can be categorized into gestures, primitive actions, interactions and group activities [2]. Gestures are the simple movements of human body parts, while 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.

One of the main problems for action recognition systems is the view-point variation during training and testing. In real time systems for action recognition, a person is typically observed from a camera having a random viewing angle; hence action recognition must be robust against varying viewing angle otherwise the accuracy of the systems decreases. Some effective methods have been proposed to solve this

problem, e.g. using multiple view sequences captured by multiple cameras then combining these sequences as training data.

Many of the recent algorithms are based on a single-view which assumes the same viewing angles during both training and testing. The performance of these systems decreases if an action video from an arbitrary view is used for testing. This drawback can be overcome by using action videos captured from multiple cameras having varying angles. In this way, 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. To overcome this issue, this work focuses on effective representation of multi-view action videos which have low dimensions hence need low storage as well as low processing requirements.

MuHAVi-uncut is a new dataset containing 17 actions (details are given in Table 1) performed different times by 7 persons. 8 CCTV cameras are used to capture execution of actions; these cameras are mounted at 45^{o} apart at the four corners and centres of rectangular action region and the recordings are continuous (of up to 3 hours) and contain the acted actions, but also the gaps and breaks in between. The interesting aspect of this dataset is that it has been obtained using a good but far from perfect foreground/shadow/background separation algorithm and thus it contains segmentation errors likely to affect higher level processes which therefore need to be robust to such errors. This dataset has an annotation at the level of action/actor/view. The time sequence for each action/actor/view contains the actor doing the action 3 to 4 times (1 time for 'ClimbLadder', DrawGraffiti' and 6 times for 'JumpOverGap'). We further manually segmented each time sequence into samples in which an actor performed an action one time and the length of each action varies with each video segment in the dataset. Data captured by Camera8 is present for 'Kick', 'PickupThroughObject', 'PullHeavyObject', 'Punch', 'RunStop' and 'WalkTurnBack' and missing for all other actions in MuHAVi-uncut dataset. The noise removal method is outlined described in section 3.1.

Table 1 Total number of samples for each action class

Action Name	Number of sequences	Action Name	Number of sequences
ClimbLadder	49	PullHeavyObject	224
CrawlOnKnees	196	Punch	224
DrawGraffiti	49	RunStop	224
DrunkWalk	105	ShotGunCollapse	140
JumpOverFence	196	SmashObject	147

JumpOverGap	294	WalkFall	147	58
Kick	224	WalkTurnBack	168	59
LookInCar	147	WaveArms	196	60
PickupThrowObject	168	**Total sequences**	**2898**	61

In this paper, a new silhouette-based multi-view human action recognition method is proposed. First of all, a noise removal step is performed to remove noise, incurred due to imperfect segmentation, from the readily available silhouette images. To overcome the high dimensionality issue incurred due to multi-camera action representations, this paper then focuses on a low-dimensional representation of multi-view data based on 2D motion templates (as is common in the literature we use the term "motion" to indicate "foreground"). For action representation Motion History Images (MHIs) [3] for each view/action video are computed, which encode how recently motion (foreground) occurred at a pixel. Moreover, for an efficient description of MHIs, Histograms of Oriented Gradients (HOG) [4] are then employed. A Nearest Neighbour (NN) classifier is used for action classification. The proposed method is evaluated on the challenging and large dataset MuHAVi-uncut [5] and outperforms similar state-of-the-art approaches. This work focuses on the recognition of a set of multi-view low-level gestures and actions; therefore this task is called multi-view action recognition.

It is true that other researchers [6] and [7] have worked with MHIs for HAR. Nevertheless, much of that type of work used "good" silhouettes tending to ignore the difficulty of segmenting people well, assuming that segmentation is a solved problem. This is understandable, but in a real system the pre-processing stages are quite critical. Therefore, we suspect that as the quality of the input decreases, many apparently good HAR algorithms will show significant worsening in performance because robustness to this type of noise appears not to have been sufficiently studied. Conversely, the use of data for which good silhouettes can be obtained (e.g. the IXMAS dataset), has led some researchers to feel (incorrectly, in our opinion) that HAR using silhouettes is not an important challenge. This probably resulted in a gradual move away from silhouettes, whereas real-world silhouettes present a real challenge. So, we have specifically built the MuHAVi-uncut dataset to see how the HAR community can address the issue that in the real world we always need to deal with "imperfect" data [8]. As an example, Fig. 1a shows the type of "perfect" silhouette (from manual segmentation of MuHAVi images) that has been used e.g. in [9] while Fig. 1b shows what state-of-the-art [8] segmentation can achieve. HAR methods based on contour-based parameters ([9], [10]), for example, would have significant difficulties with such data. Thus we propose here to deal with quite noisy silhouettes, evaluating the use of pre-filtering and HOG as a means for

dealing with such noise and comparing it with competitive methods that also use silhouettes. It would be interesting, but out of the scope of this paper given the difficulties in re-implementing algorithms reported in the literature, to see how other type of HAR methods (e.g. based on CNNs) cope with the noisy

(a) (b)

MuHAVi-uncut data.

Fig. 1: Silhouette images of 'Walk' action taken from

a MHAVi-MAS and
b MuHaVi-uncut dataset

The rest of the paper is organized as follows: Section 2 presents a review of single and multi-view human action recognition approaches. Section 3 describes the proposed approach in detail. Section 4 presents an explanation of experimental results on the benchmark dataset. Finally, Section 5 concludes with some discussions and ideas for future work.

2. Related Work

Extensive computer vision approaches and methodologies have been presented in the literature in the field of action recognition. Since action recognition in video sequences is a challenging problem, many methods have deliberately simplified settings. Action recognition methods are characterized depending on the visual information used for action representation description. Therefore, existing approaches can be divided according to being single or multi-view action recognition.

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 found in the literature. These are holistic methods, local feature methods, and geometric human body features. Holistic approaches uses shape [11] or motion based [12] information. Shape based methods do

not depend upon moving person's clothes color, texture, and luminance therefore these are potentially useful for action representation. Motion based approaches are unpredictable in case of motion discontinuities, low quality videos and small variation in background. In geometric human body features, action recognition is done on the basis on locating body parts along with movements [13]. Local space-time features or interest points [14] compute the shape and motion characteristics in video locally and a feature descriptor is used [15] to describe these features efficiently. These features need to be robust against scale, background motion and clutter. Such local features do not require object segmentation; hence they are extracted directly from video frames.

As single-view approaches use one camera for capturing human action, these approaches usually need to have the same or a similar camera-view during training and testing. If this condition is not met, the accuracy of these approaches may significantly decrease because same actions look quite different when captured by arbitrary viewing angles [16]. Therefore view-invariant action recognition has attracted attention in the last decade by exploiting multi-camera for better action description, as surveyed in [17], [18].

Multi-view approaches may be divided in two categories [19]: 3D and 2D multi-view methods. In most of 3D methods, 2D human body silhouettes are joined to have a "3D human body pose" representation and actions descriptions are obtained by a series of these successive "3D human body poses". Examples of "3D human body poses" representations are visual hulls, motion history volumes [20], optical flow [21], spatio-temporal volumes [22] and multi-view postures [23]. These methods tend to require a (fixed) multi-camera setup during training as well as testing stages. The 3D human body representation based on sequences of key poses as introduced by [9] in which pose representation is based on contour points of human silhouettes. They used a feature fusion approach to fuse multi-view features. 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 the actor is outside the range of camera view or due to occlusion [24].

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 in a network. View-invariant action representations is proposed in [25], [26], [27] and classification is carried out by training a universal classifier for all available views or by using multiple classifiers for training [28]. The final result is obtained by fusing the results from all the classifiers.

The second direction is based on multi-view learning during training and testing of unknown action is done based on these learned features. In [29], [10] features are learned individually from each camera-view independently of other camera-views. As no feature fusion is used in this approach therefore there is no need to have all camera views available during training stage. The advantage of this type of approach is that it can handle missing views of an action. The third direction is based on cross-view action recognition. In this approach classes in one view are used for training while others for testing. Many techniques have been anticipated using cross-view action recognition such as transfer learning [30], [31], information maximization [32], etc.

3. Proposed Method

The goal of this work is to have an efficient and low dimensional representation of large multi-view action classes. The proposed multi-view human action recognition method is divided into four components: noise removal, features extraction using MHIs, features description based on HOG descriptor and action recognition. A detail of the multi-camera human action recognition framework is presented in Fig. 2.

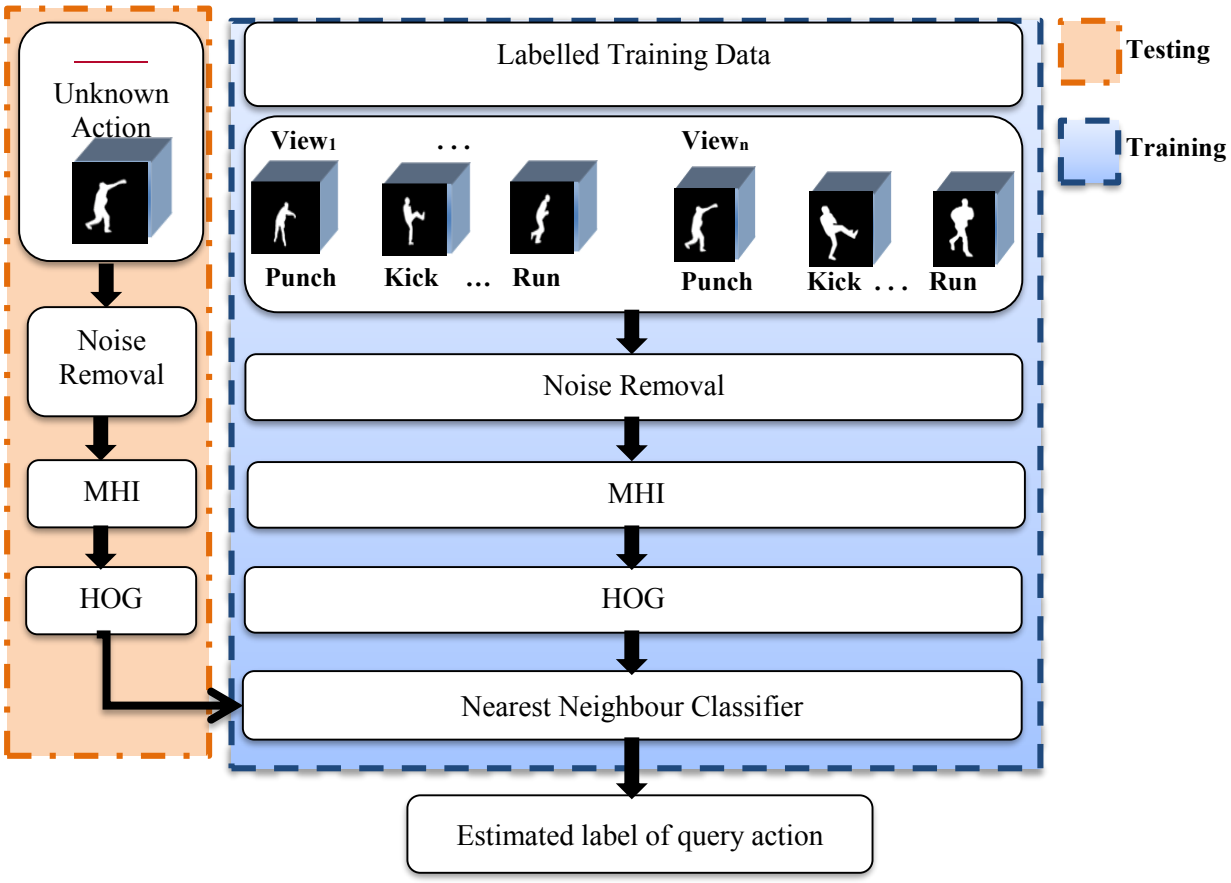


Fig. 2*. General overview of the proposed multi-view human action recognition framework*

It is assumed that the input video sequences have been processed by an algorithm that separates foreground from background [33], [8]. It is also recognised that even a state-of-the-art background removal method will produce noise, especially if it operates so as not to miss foreground regions. Therefore, we first remove noise from every action video and then MHIs are calculated from the resulting silhouettes. The resulting MHIs are centred with respect to the centre of gravity of each detected object so as to have location independent representation. Then the resulting MHIs are scaled to a fixed sized to have a scale invariant representation. Finally, the HOG descriptor is computed for each MHI. The Nearest-Neighbour classifier is trained with HOG descriptions of the MHIs. In the following section, the components of the proposed framework are presented in more detail.

3.1. Noise removal

In our method, silhouette images are directly used; therefore pre-processing steps are necessary to remove noise and shadows incurred due to imperfect segmentation. The MuHAVi-uncut dataset has many challenges in which the major one is that of lighting complexity. This dataset is captured in conditions of non-uniform background, fluctuating harsh illumination, and cast and self-shadows. Shadow points may be confused with foreground as they are difficult to model as part of the background and they move with the object and hence can reduce the performance of object detection, recognition, etc.

Silhouette data for MuHAVi known as MuHAVi-uncut was obtained using the background subtraction method called Self-Adaptive Gaussian Mixture Model (SAGMM) used in [33] proposed by [34], and available from the corresponding author of this paper. The algorithm has a dynamically adaptive learning rate and models global illumination changes of the background frame by frame. However, because of the sometimes harsh illumination the data has noise and shadows as shown in Fig. 3a. The MuHAVi-uncut silhouette images have three types of noises;

i. small blobs e.g. salt and pepper noise,
ii. shadows,
iii. moving objects in the background

First the salt and pepper noise has been removed using a median filter of size depending on the size of noise (in our case we used a 15 x 15 sized filter because the salt and pepper noise in the MuHAVi-uncut has a size of about 15x15 pixels). Median filters are widely used for removing salt and pepper noise as it effectively removes noise while preserving edges. Sometimes the filter size has to vary depending upon the size of noise in the images.

Secondly moving background objects have been removed by finding the size of every blob in the silhouette image. As the background objects are at a distance therefore their size is smaller than the foreground actor hence the blobs having larger areas are considered as foreground and all other objects are considered as background and removed. As mentioned earlier, the MuHAVi-uncut dataset has been segmented by SAGMM algorithm [8] that codes the output images into three levels: 255 for Foreground, 0 for Background and 127 for possible shadows, so it is relatively simple to remove the potential shadows by a fixed threshold (T=128) as indicated by Equation 1. After noise removal from Fig. 3a, a typical resulting image is shown in Fig. 3b. Finally the foreground is further refined to fill the holes using region filling operation based on 4 or 8 connected neighborhoods.

$$\mathbf{I} = \begin{cases} \mathbf{255} & \mathbf{if\ f(x,y)>}\boldsymbol{T} \\ \mathbf{0} & \mathbf{otherwise} \end{cases} \tag{1}$$

a *b*

Fig. 3. *Illustration of noise removal of silhouette image of MuHAVi dataset* a Before noise removal b After noise removal

3.2. Feature extraction

The input features used in this work are based on MHI masks or templates. 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 provides information of motion (foreground) 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 indicates that there exists motion (foreground) at time *t* at location (*x, y*). The MHI at time *t* is computed as:

$$\mathbf{MHI_{\tau}} = \begin{cases} \boldsymbol{\tau} & \mathbf{if\ I(x,y,t)=1} \\ \mathbf{max(0,MHI_{t-1}(x,y)-1)} & \mathbf{otherwise} \end{cases} \tag{2}$$

where τ is the number of frames used for computing the MHI template. Automatic temporal segmentation is outside the scope of this work therefore we have assumed to have pre-segmented action sequences and the data about temporal segmentation is available on the dataset. In all experiments, the value of τ is taken to be equal to the total number of frames in an action video and τ may be different for different actions. The resulting MHI can represent the foreground sequence in a compact manner. The silhouette sequence belonging to one action is compressed into a gray scale image, where most recent motion is represented by brighter gray-value pixels as shown in Fig.4, preserving 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 scale and location invariant representation. Illumination and contrast invariance representation can be achieved by dividing each pixel in the MHI by the sum of the total pixels of $\boldsymbol{MHI\tau}$ to have unit sum as given under:

$$\boldsymbol{MHI_n} = \frac{\boldsymbol{MHI\tau}}{\sum_{x=1}^{M}\sum_{y=1}^{N}\boldsymbol{MHI\tau}(x,y)} \tag{3}$$

Where M is the total number of rows and N is the total number of columns in $\boldsymbol{MHI\tau}$ and $\boldsymbol{MHI_n}$ is the normalized version of the original $\boldsymbol{MHI\tau}$ as obtained from Equation 2.

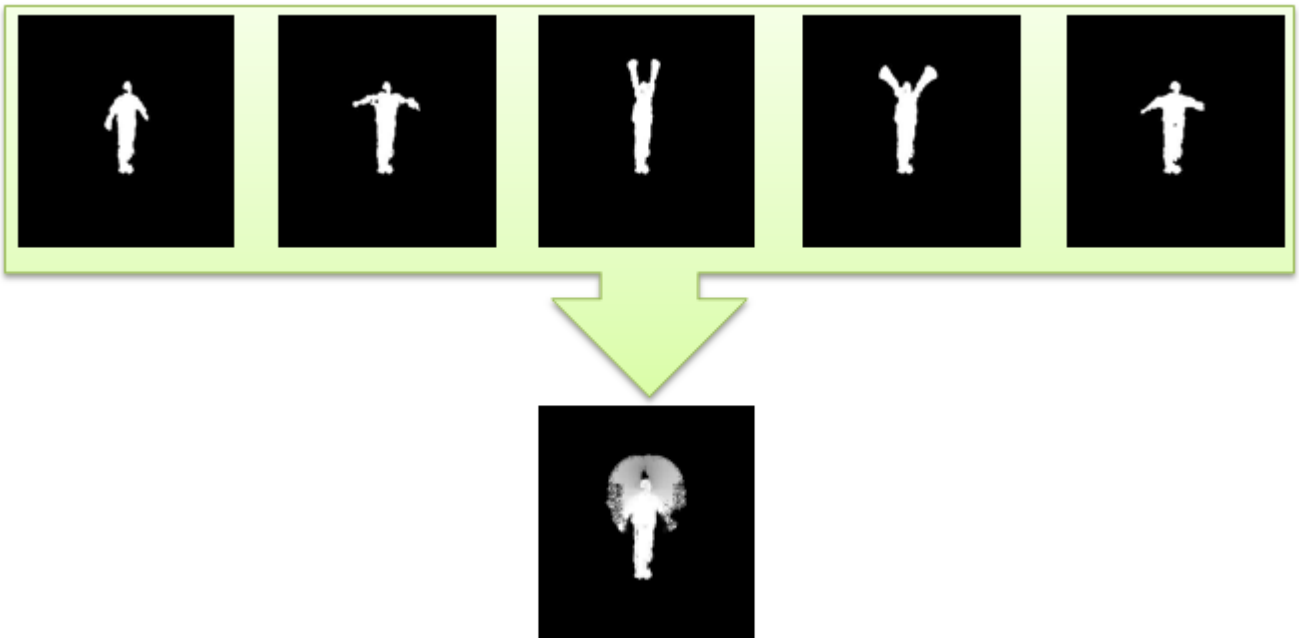


Fig.4. *MHI of action "WaveArms" from MuHAVi dataset*

View-point dependencies can be removed by generating a 3D model of the human body as in [20]. However, this method would require many fixed calibrated cameras during training as well as testing phase, which is not realistic in real time scenarios. To have view-independent recognition we have used an exhaustive search method [35] in which multiple cameras are used during training phase to obtain HOG-

MHIs for each view independently. During recognition, the test sequence from an arbitrary view will be compared with all of the learned actions and the best matching sequence is identified as in [36] and [37]. Our approach does not need calibrated cameras or feature fusion of multiple views; hence, one camera is used for testing. In many real world applications, subjects are observed by a few (and many times just a single) cameras due to constraints like occlusion, view point etc., therefore our approach is suitable for such scenarios. Although it is true that action classification performance may be improved by simultaneously using multiple views when deciding how to classify a previously unseen sequence, in the real world of public space surveillance it is extremely rare to find multiple views of the same people/action and therefore we have assumed that a decision needs to be made using a single view. Nevertheless, we show here how the multiple views are used to train the system that even though it is simple, it manages to do better than comparable approaches. A more sophisticated way to include multiple views e.g. using manifolds [38], can be used but the paper would be too lengthy to include such work. The simultaneous use of multiple views may be result in accuracy improvements but it will also affect computational complexity, something that is quite relevant for real-time processing.

MHI representation is less sensitive to noise, shadows, missing body parts and holes [39] in the foreground objects resulting from imperfect segmentation. The MHI is sensitive to direction of motion of action; therefore, it may discriminate between actions of opposite directions (e.g. run left and run right). The MHI can encode a range of times in a single matrix which makes this method more computationally efficient.

3.3. Feature descriptor

In this work HOG based features are used to describe MHIs. HOG feature descriptors [4] 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 [29]). The implementation of the HOG descriptor is based on the default descriptor suggested in [4] with the following attributes: grayscale with no gamma correction; [-1 0 1] gradient filter with no smoothing; 8x8 pixel blocks of four 4x4 pixel cells (default is block is of size 16x16 pixel having four 8x8 pixel cells); linear gradient voting into 8 orientation bins on the interval of 0°-180°; non-overlapping block (default is block overlap of half of the block size); L2-Norm block normalization. As one block has 4 cells therefore for each block these 4 histograms (having 8 orientation bins) are concatenated into a row vector having 32 dimensions. The final descriptor is formed by concatenating the histograms of all blocks into a row vector of size 32 multiplied by total number of

blocks in an image. We have divided images of 100x50 pixels into blocks of 8x8 pixels therefore we have 78 blocks (13 in vertical and 6 in vertical direction) per image. For one image the final descriptor has 2496 dimensions.

3.4. Action recognition

To obtain a baseline which future researchers can improve upon, we use a simple yet effective Nearest-Neighbour classifier approach to obtain a nearest class label which was originally suggested by [40]. Owing to its simplicity it is used in a large number of classification problems [41] and it does not require any prior assumptions about the distributions of the training examples of data, similarly this can efficiently classify between large action classes [40]. For training, MHI samples and their corresponding HOG-based description is computed and the NN classifier is trained for all training sequences. For a given test MHI sample, its corresponding HOG descriptor is calculated and given to the classifier, which will find the most similar sample (under some distance measure) from the labelled training data so as to assign a label to the test sample. The accuracy rates of this classifier rely significantly on the distance measure that is used. Different types of distance metrics are reported for this classifier, e.g. Cityblock, Chebychev, Euclidean and Minkowski. Euclidean metric is used in our work to measure distances between query and the training samples [using (4)] because it is the most commonly used distance for action recognition [10]. According to Marinaki et al. [42], the Euclidean metric produces very significant and promising results in terms of speed and accuracy and this can be applied to many difficult problems. A valuable feature of the Euclidean distance is that it is preserved under orthonormal transforms [43]. Another reason for selecting Euclidean distance is that, as we are using noisy MuHAVi silhouette dataset, if signals are corrupted by noise it has been found to be an appropriate distance measure forestimation [44]. We have also compared different types of distance metrics and found that Euclidean metric is better in terms of speed and accuracy than other metrics, but because of space constraints we are unable to include those results here

$$d = \sqrt{\sum_{j=1}^{n} (p_j - q_j)^2} \tag{4}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are the HOG vectors for test and training samples respectively and n is the length of the vector.

4. Experimental Results

The efficiency of algorithm is tested on the MuHAVi-uncut dataset [5]. Some of the examples of our proposed method are shown in Fig.5. In the following section, validation schemes used for the proposed

algorithm are explained. For all of the experiments in this section the value of τ is set to be equal to the total number of frames in an action, 8 orientation bins and an 8x8 block size for the HOG descriptor is used. All experiments are performed using MATLAB 7 with Intel core i3 at 1.70 GHz, 4GB RAM, 64 bit operating system. Three types of cross validation schemes are used for performance evaluation: leave-one-sequence-out (LOSO), leave-one-actor-out (LOAO) and leave-one-camera-out (LOCO).

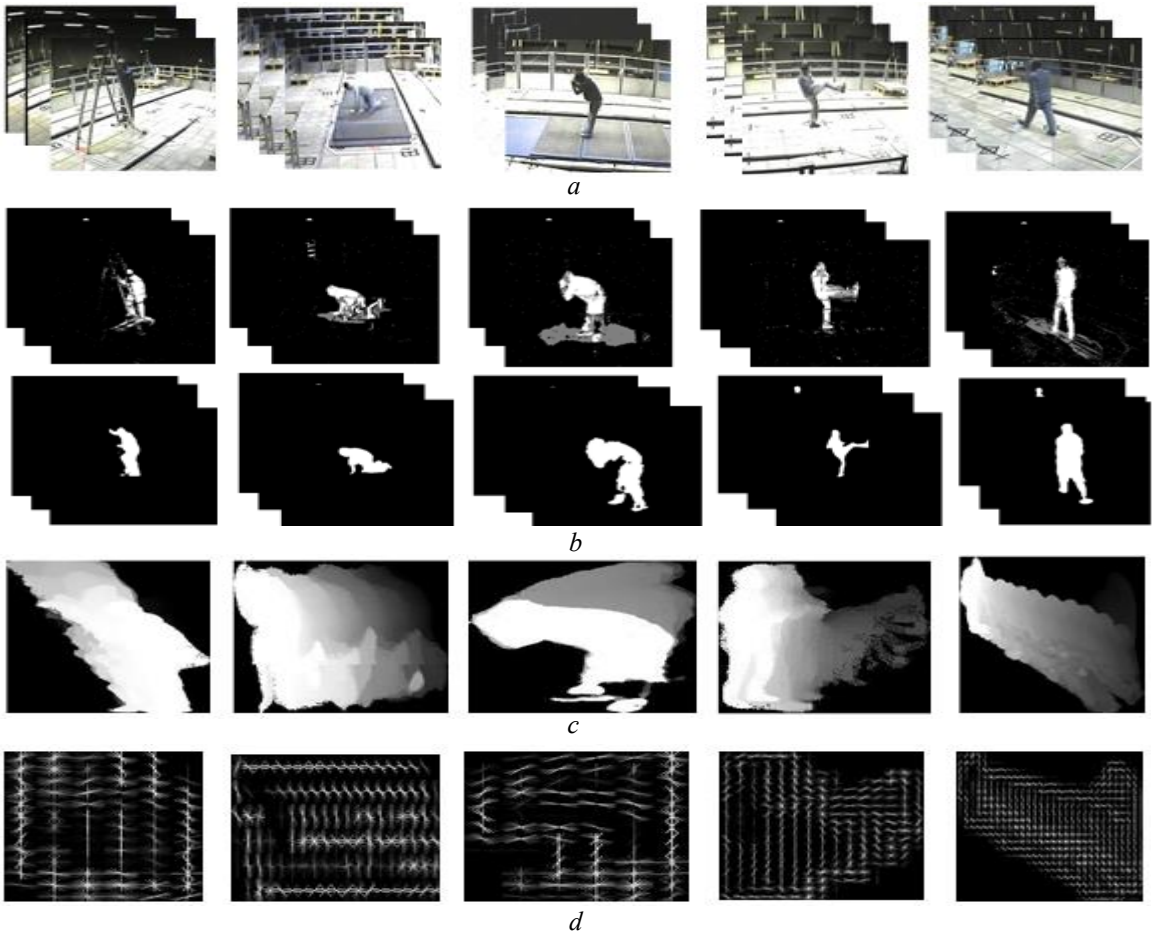


Fig.5. *Illustration of the proposed work on 5 actions from MuHAVi-uncut*

a RGB sequences of "ClimbLadder", "CrawlOnKnees", "LookInCar", "Kick" and "WalkTurnBack" actions from left to right

b Row1: noisy silhouettes and Row2: noise less silhouettes after noise removal

c MHI based representation

d HOG description of MHIs

4.1. Cross validation schemes

4.1.1 Leave-one-sequence-out (LOSO): In LOSO, the Nearest-Neighbour classifier is trained on all sequences except one and that left-one is used for testing. The process is repeated for all such combinations and the mean and standard deviation of accuracies are computed to assess performance. The proposed method attains very promising accuracy rates of 95.9% for MuHAVi-uncut with a low standard

deviation of 3 among action classes and the resultant confusion matrix is shown in Fig. 6. The results show that "DrawGraffiti" has less accuracy compared to other actions and it is confused with "LookInCar" and "SmashObject", "PickupthroughObject", "Punch" and "SmashObject" actions due to the similarity in their poses. Similarly, "DrunkWalk" is misclassified with "PickupthroughObject" "WalkTurnBack" and "SmashObject" which shows that a HOG based description of MHIs is not discriminative enough. "JumpOverGap" action has a very short duration (about 30 to 40 frames) therefore an MHI based representation is not suitable. Nevertheless, the results indicate that our proposed algorithm is robust and efficient in circumstances where an action is made up of multiple primitive actions e.g. "Walk" and "Fall" combined to make single "WalkFall" action. The comparison of our proposed method with similar state-of-the-art approaches are presented later in section 4.2.

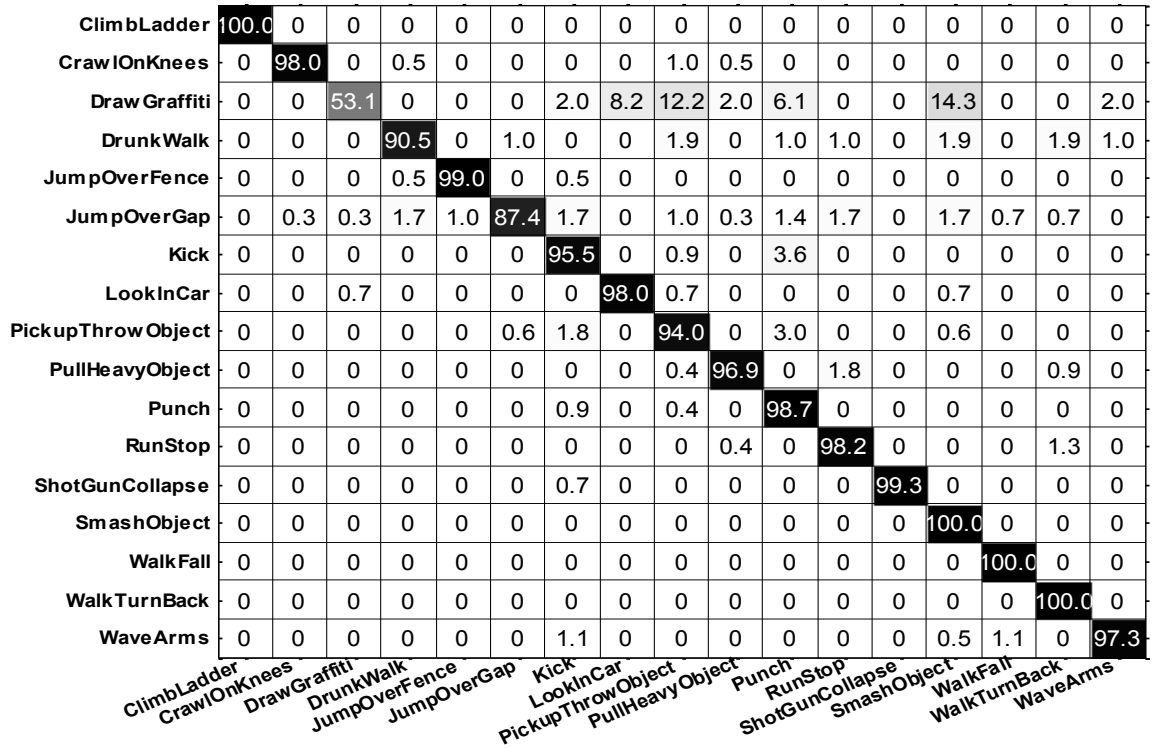


Fig. 6. *Confusion matrix for MuHAVi using LOSO (Average accuracy=95.9% and standard deviation=3)*

4.1.2 Leave-one-actor-out (LOAO): MuHAVi-uncut dataset contains seven actors. In the LOAO cross validation scheme the experiment is performed by training the classifier on all sequences of all actors except one, which is then used for testing. The average accuracy is calculated by alternatively testing all actors. "DrawGraffiti" again shows less accuracy and it confused with "PickupthroughObject", "JumpOverGap", "Kick" and "SmashObject" actions due to similarity in their poses. As "PickupThroughObject" and "Punch" actions have greater similarity in their poses therefore these are confused with each other as shown in the confusion matrix shown in Fig. 7. Similarly "Punch" and "Kick"

actions are confused with each other both actions has similar "Guard" poses before action starts. Under LOAO our approach is more robust as compared to other state-of-the-art approaches, as discussed in section 4.2. This experiment got an accuracy rate of 84.1% on MuHAVi-uncut dataset with a low standard deviation of 4 among action classes.

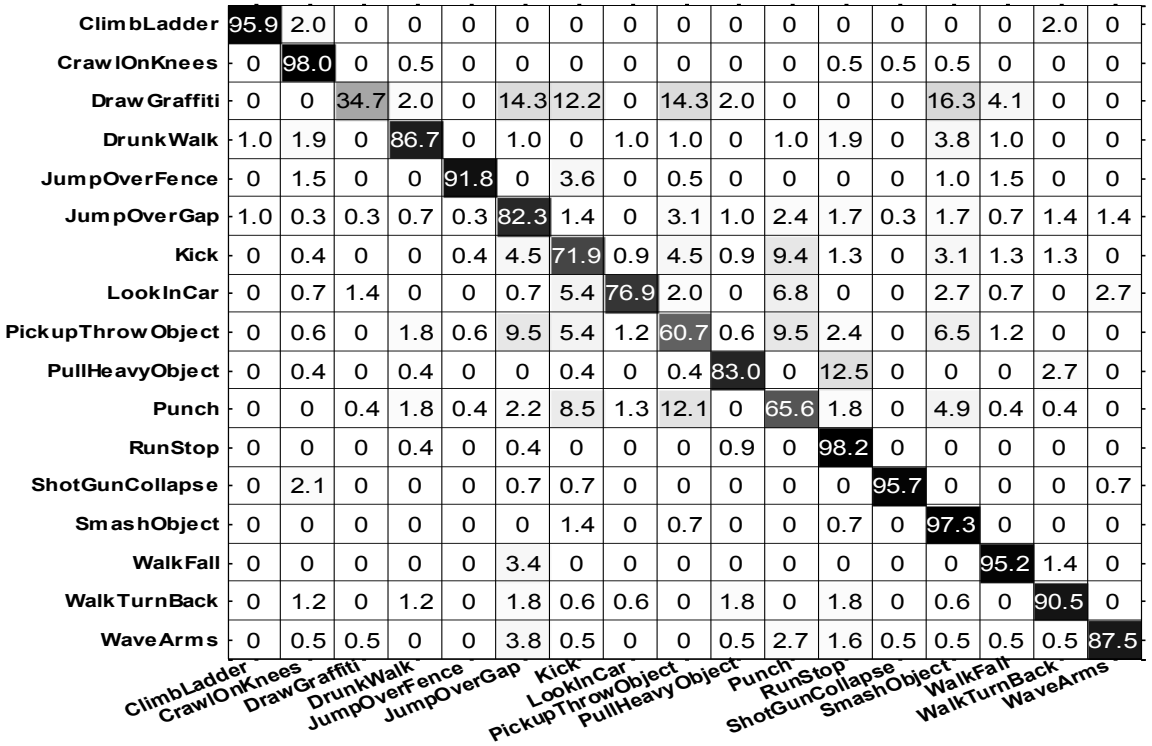


Fig. *7. Confusion matrix for MuHAVi using LOAO (Average accuracy=84.1%* and standard deviation=4*)*

4.1.3 Leave-one-camera (LOCO): In the LOCO cross validation scheme the classifier is trained on all sequences from all cameras except one, which is then used for testing. MuHAVi-uncut contains 8 camera-views available therefore average accuracy is calculated by testing alternatively on all of the available camera-views. This validation scheme tests robustness of a proposed algorithm to changes in viewpoint. This experiment got accuracy rate of 52.2% with a low standard deviation of 5 among action classes and the resultant confusion matrix is shown in Fig. 8. In particular, we notice a low performance of our approach for actions where there is high self-occlusion in some camera-views e.g. "WalkFall", "DrawGraffiti" and "Punch". The low accuracy rate using LOCO indicates that view-independence is a challenge for future workers. The comparison of our proposed method with similar state-of-the-art approaches are presented later in section 4.2.

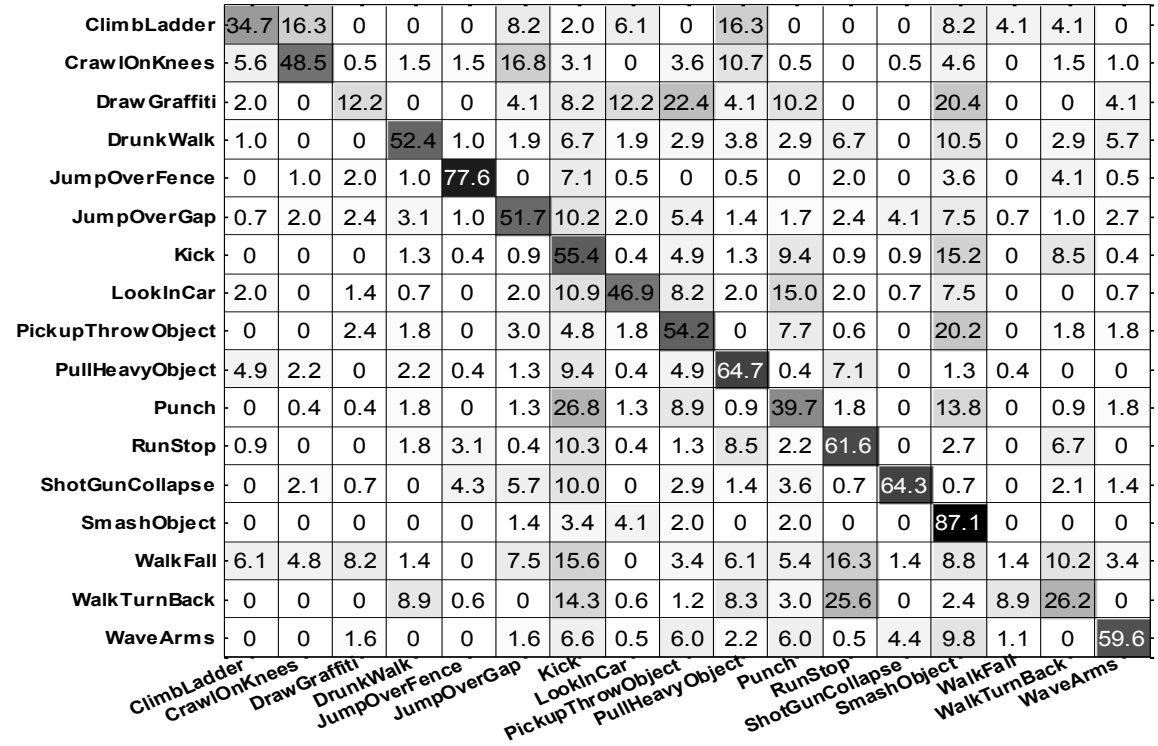
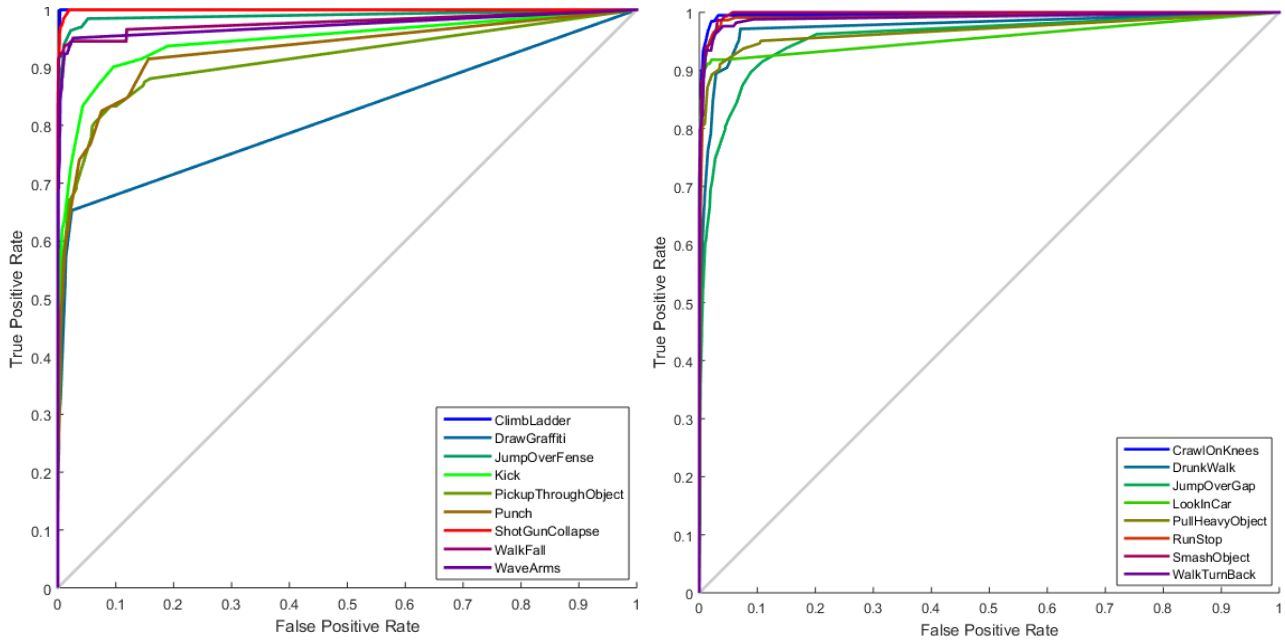


Fig. 8 *Confusion matrix for MuHAVi MHI using LOCO (Average accuracy=52.2%* and standard deviation=5*)*

4.2. Performance curves

To show the classifier (KNN) performance we calculated the ROC curves, based on LOAO based cross validation technique, for 17 MuHAVi actions (shown in Fig. 9 ROC Curves for seventeen actions. The dotted diagonal line *shows the random prediction.*). The ROC curves are calculated using one versus all classes and instead of using separate plots for each class we plotted 9 classes in Fig. 9a and remaining 8 classes Fig. 9b, due to space limitation. The area under the curve for "ClimbLadder", "CrawlOnKnees", "DrwaGraffiti", "DrunkWalk", "JumpOverFence", "JumpOverGap", "Kick", "LookInCar", "PickupThroughObject", "PullHeavyObject", "Punch", "RunStop", "ShotGunCollapse", "SmashObject", "WalkFall", "WalkTurnBack" and "WaveArms" is 0.999, 0.990, 0.809, 0.961, 0.984, 0.973, 0.914, 0.990, 0.904, 0.989, 0.921, 0.994, 0.996, 0.998, 0.971, 0.992 and 0.970 respectively. The area under the ROC curves, shown in Fig. 9, shows that our approach is able to correctly classify most of the actions. Fig. 9 also shows that "WalkTurnBAck" and "RunStop" are better classified than "Punch" and "PickupThroughObject". As discussed earlier in Section 4.1.1, this is because "Punch" and "PickupThroughObject have similar poses therefore it is comparatively difficult to differentiate between them, but our method is still able to differentiate between these to some extent. Similarly the area under the ROC curve for "DrawGraffiti" is smaller than for other actions because it has high similarity in poses with "LookInCar" and "SmashObject", "PickupthroughObject", "Punch" and "SmashObject" actions.

15



a *b*

Fig. 9 *ROC Curves for seventeen actions. The dotted diagonal line shows the random prediction.*

a ROC curves for ClimbLadder, DrawGraffiti, JumpOverFence, Kick, PickupThroughObject, Punch, ShotGunCollapse, WalkFall,WaveArms
b ROC curves for CrawlOnKnees, DrunkWalk, JumpOverGap, LookInCar, PullHeavyObjects, RunStop, SmashObjec, WalkTurnBack

4.3. Comparison of results

In this paper we have focused on dealing with fairly noisy silhouettes, proposing the use of pre-filtering and HOG as a means for dealing with such noise and comparing it with competing methods ([9], [29] and [10]) that also use silhouettes. Our method might at first sight appear to be similar to differential MHI with HOG used in [6], but their method does not use silhouettes but interframe RGB differences, more similar to gait energy information (GEI) used in [6]. They have reported results on the KTH dataset [45], which in terms of noise is a fairly simple dataset. Similarly work presented in [46] also used MHI and HOG but for gesture emotion recognition. Recently authors [47] used patterns of motion and histogram of oriented gradient for pedestrian activity classification. Their "actions" have to do with direction of walking. They used inter-frame subtraction thus avoiding silhouettes. Thus they have focused entirely on the variation of action direction and action type excluding the other factors like camera movement, illumination and background variation. The video clips are taken in an indoor environment with a constant and stable background and illumination situation with a Nikon D3200 camera. They used their own simple dataset and they avoid background/illumination variations.

The comparison of the proposed method on the MuHAVi-uncut with many other human action recognition approaches is difficult because, as far as we know, no one has reported on this new dataset before. Therefore we have re-implemented the approach of [10] and have access to the software of [9] and [29] for comparisons. [9] and [10] have used LOSO, LOAO and LOCO based validation techniques on different datasets whereas [29] only used LOAO based validation technique and have reported good results on other datasets (MuHAVi-14, MuHAVi-8, IXMAS etc.). For comparison against other methods ([9], [29] and [10]) we have applied the same noise reduction process for all methods and then evaluated each algorithm using the cleaned silhouettes sequences to ensure that we are comparing like with like.

First, the proposed method is compared with [29]; which used MHIs as 2D motion template on their own, whereas we have used HOG based description of MHIs. They have also considered temporal segmentation which is out of the scope of our work. They have computed multiple MHIs from a video, and τ i.e. the number of images used for generating the MHI is small (10 frames) whereas we have used τ = total frames in an action video which made our method more computationally efficient. By using simple nearest neighbour classifier we obtained a high accuracy rate (84.1%) as compared to complex Hidden Markov Model (HMM) used by [29] based on LOAO cross validation scheme (shown in Table 2).

Secondly, the proposed method is compared with [9] and [10] because they have used different cross validation schemes and also have reported computational complexity in their work so comparison with their work became possible (although we had to implement their method as source code was not available). Contour-based pose feature from silhouettes are used in [9] and [10] and features are clustered to build a set of prototypical key poses. As compared to their work we have extracted MHI/HOG features and shown that a single MHI for an action video is computationally efficient, whereas extraction of contour based features from every frame of a video is computationally expensive and problematic for real time operation. MuHAVi-uncut has noisy silhouettes that do not have sharp contours and even with noise and shadow removal, contour based features are affected by this more realistic dataset. Even without using feature fusion from multiple views, as [9] does, we still obtained high accuracy rates. The proposed method outperforms [9] and [10] using three types of cross validation schemes as illustrated in Table 2, achieving accuracy improvements 10%, 3% and 2% higher than [9] for LOSO, LOAO, LOCO validation schemes respectively. Similarly, the proposed approach achieved the accuracy 29%, 28% and 21% higher than [10] for LOSO, LOAO, LOCO validation schemes respectively.

Table 2 Comparison results on the basis of recognition accuracy

Approach	MuHAVi-uncut		Cross-validation scheme
	Accuracy (%)	Standard Deviation	
[9]	86.5	8	LOSO
[10]	56.6	8.4	
This paper	**95.9**	**3**	
[9]	81.5	5	LOAO
[29]	83.9	6	
[10]	56.7	11	
This paper	**84.1**	**4**	
[9]	50.4	5.3	LOCO
[10]	31.4	6	
This paper	**52.2**	**5**	

4.4 Computational complexity

The algorithms compared here were all implemented in the same environment (MATLAB) and the same computer, so computational complexity differences can be estimated by measuring times and frame rates. In MuHAVi-uncut, there is a total of 17 action classes having 2898 sequences which are further composed of total 330,378 frames and each frame has resolution of 720 × 576 pixels. Using the LOAO validation scheme, training and testing of 2898 sequences takes 3992 s, i.e. on average it takes 1.38 s per sequence at a rate of 82.8 frames/s (FPS). This test includes all the steps of computing MHIs, HOG description and the final recognition process using NN classifier, and assuming that cleaned silhouette sequences are available for all experiments.

For MuHAVi-uncut, Chaaraoui et al. [9] achieved 9.1 FPS, which shows that our proposed approach is about nine times faster than their method. Similarly, Orrite et al. [29] achieved only 2 FPS, 40 times slower than our method, which limits their method for real-time action recognition of human actions. The frame rate achieved by Cheema et al. [10] is 9.9 FPS, i.e. our method is about eight times faster. Table 3 shows a comparison of accuracy and recognition speeds based on the LOAO validation scheme. As compared with [9, 29, 10], our algorithm takes less time because we have only need to extract MHI/HOG features which shows that the use of a single MHI for an action video is computationally efficient, as is clear in Table 3, for example.

Table 3 Comparison results on the basis of recognition accuracy and speed

Approach	Accuracy (%)	FPS
[9]	81.5	9.1
[29]	83.9	1.7
[10]	56.7	9.9
This paper	**84.1**	**82.8**

5. Conclusion and Future Work

In this research work, a silhouette-based human action recognition method has been proposed for a multi-camera dataset. A HOG based description of MHIs has resulted in high accuracy rates using a simple Nearest-Neighbour classifier as compared to similar state-of-art methods. With MuHAVi-uncut, as far as we know, these are the first results provided with this noisy and large dataset and as such we have opted for Nearest-Neighbour for simplicity. Probably the use of more sophisticated classifiers (such as SVMs or manifold-based schemes) would improve these results, but the relatively high accuracy obtained using the simpler Nearest-Neighbour approach supports the appropriateness of choosing HOG. 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 [29]).

The proposed method can cope with high-dimensionally data incurred due to multi-camera dataset, which is achieved by using single MHI for whole image sequence of each action. The proposed method does not require feature fusion; therefore, it is suitable for real-time scenarios because the subjects are not necessarily visible from all camera-views.

The performance of the algorithm has been tested on a large number of noisy silhouettes i.e. MuHAVi-uncut. HOG based description of MHIs has provided an efficient description of action classes in spite of the fact that MuHAVi-uncut has noisy silhouettes. The proposed method achieved a high accuracy rate 95.4% on MuHAVi-uncut using LOSO validation. A frame rate of 82.8 FPS is achieved; hence it 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 future work, clustering techniques can be used to obtain key MHI poses per action class, which may describes an action with more distinctive representation. Along with clustering techniques, feature descriptors other than HOG descriptor can be used. Along with different feature descriptors, more sophisticated classifiers (such as SVMs or manifold-based schemes) would be used to improve these results. The problems of automatic temporal segmentation and above all that of view independence remain important challenges for the future.

6. Acknowledgements

Sergio A Velastin acknowledges the Chilean National Science and Technology Council (CONICYT) for its funding under grant CONICYT-Fondecyt Regular no. 1140209 (“OBSERVE”). He is currently funded by 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.

7. References

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

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

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

[4]. Dalal, N., Triggs, B.: 'Histograms of Oriented Gradients for Human Detection', In *CVPR*, 2005, pp. 886-893.

[5]. Singh, S., Velastin, S.A., Ragheb, H.: 'MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods', Advanced Video and Signal Based Surveillance (AVSS), Seventh IEEE International Conference on , 2010, pp. 48,55.

[6]. Wu, Di,, Shao, L.: 'Silhouette analysis-based action recognition via exploiting human poses', Circuits and Systems for Video Technology, IEEE Transactions on 23.2, pp. 236-243, 2013.

[7]. Huang, C.P., Hsieh, C.H., Lai, K.T., Huang, W.Y.: 'Human action recognition using histogram of oriented gradient of motion history image', Instrumentation, Measurement, Computer, Communication and Control, 2011 First International Conference on. IEEE, 2011.

[8]. Sepulveda, J., Velastin, S.A.: 'F1 Score Assesment of Gaussian Mixture Background Subtraction Algorithms Using the MuHAVi Dataset', 6th IET International Conference on Imaging for Crime Detection and Prevention (ICDP-15), 15-17 July, London (2015).

[9]. Chaaraoui, A. A., Climent-Pérez, P., Flórez-Revuelta, F.: 'Silhouette-based human action recognition using sequences of key poses', Pattern Recognition Letters, Volume 34, Issue 15, 1 November 2013, pp. 1799-1807.

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

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

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

[13]. Ramanan, D., Forsyth, D.A., Zisserman, A.: 'Tracking people by learning their appearance', IEEE T-PAMI, 2007, 29:65-81.

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

[15]. Lowe, D. G.: 'Distinctive image features from scale-invariant keypoints', International Journal of Computer Vision, 2004, 60(2):91–110.

[16]. Rudoy, D., Manor, L. Z.: 'Viewpoint Selection for Human Actions', IJCV, 2012, vol. 97, pp. 243–25.

[17]. Ji, X., Liu, H.: 'Advances in view-invariant human motion analysis: A review', Trans. Sys. Man Cyber: Part C, 2010, vol. 40, no. 1, pp. 13–24.

[18]. Holte, M. B., Moeslund, T. B., Tran, C., Trivedy, M. M.: 'Human Action Recognition using Multiple Views: A Comparative Perspective on Recent Developments', ACM HGBU, 2011, pp. 47-52.

[19]. Iosifidis, A., Tefas, A., Pitas, I.: 'Multi-view Human Action Recognition: A Survey', Intelligent Information Hiding and Multimedia Signal Processing, Ninth International Conference on , 2013, pp. 522-525, 16-18.

[20]. Weinland, D., Ronfard, R., Boyer, E.: 'Free viewpoint action recognition using motion history volumes', CVIU, 2006, vol. 104, no. 2, pp. 249–257.

[21]. Holte, M., Moeslund, T., Nikolaidis, N., Pitas, I.: '3D human action recognition for multi-view camera systems', 3DIMPVT, 2011, pp. 342-349.

[22]. Yan, P., Khan, S., Shah, M.: 'Learning 4D action feature models for arbitrary view action recognition', CVPR, 2008, pp. 1-7.

[23]. Gkalelis, N. Nikolaidis, N., Pitas, I.: 'View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation', IEEE ICME, 2009, pp. 394-397.

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

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

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

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

[28]. Ahmad, M., Lee, S. W.: ‘HMM-based human action recognition using multiview image sequences’, ICPR, 2006, pp. 263-266.

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

[30]. Li, R.: ‘Discriminative virtual views for cross-view action recognition’, CVPR, 2012, pp. 2855-2862.

[31]. Li, B., Campls, O. I., Sznaier, M.: ‘Cross-view Activity Recognition using Hankelets’, CVPR 2012, pp. 1362-1369.

[32]. Liu, J., Luo, J., Shah, M.: ‘Recognizing realistic actions from videos "in the wild"’. In CVPR, 2009, pp. 1996-2003.

[33]. Sepulveda, J., Velastin, S.A.: ‘Evaluation of Background Subtraction Algorithms using MuHAVi, a Multicamera Human Action Video Dataset’, 6th Chilean Conference on Pattern Recognition, November 10-14, Talca, Chile (2014), pp. 10-14.

[34]. Chen, Z., Ellis, T.: ‘Self-adaptive gaussian mixture model for urban traffic monitoring system’, In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pages 1769–1776, 2011.

[35]. Weinland, D., Ronfard, R., Boyer, E.: ‘A survey of vision-based methods for action representation, segmentation and recognition’, Computer Vision and Image Understanding, vol. 115, issue 2, pp. 224-241, February 2011.

[36]. Ahmad, M., Lee, S.W.: ‘Hmm-based human action recognition using multiview image sequences’, In International Conference on Pattern Recognition, vol. 1, pp. 263-266, 2006.

[37]. Ogale, A., Karapurkar, A. Guerra-Filho, G., Aloimonos, Y.: ‘View-invariant identification of pose sequences for action recognition’, In: VACE, 2004.

[38]. Lewandowski, M., Makris, D., Velastin, S.A., Nebel, J.C.: ‘Structural Laplacian Eigenmaps for modelling sets of multivariate sequences’, 'IEEE Transactions on Systems, Man and Cybernetics, Part B', DOI: 10.1109/TCYB.2013.2277664, 2014.

[39]. Ahad, Md., Tan, J., Kim, H., Ishikawa, S.: ‘Motion history image: its variants and applications’, Machine Vision and Applications, pp. 1–27, Oct. 2010.

[40]. Cover, T. M., Hart, P. E.: ‘Nearest neighbor pattern classification’, IEEE Trans. Inform. Theory, vol. IT-13, pp. 21–26, 1967.

[41]. Fayed, H., Atiya, A.: ‘A Novel Template Reduction Approach for the -Nearest Neighbor Method’, IEEE Trans on Neural Network , Vol. 20, No. 5, pp. 890-896, May 2009.

[42]. Marinaki, M., Marinakis, Y., Doumpos, M., Matsatsinis, N., & Zopounidis, C.: 'A comparison of several nearest neighbor classifier metrics using Tabu Search algorithm for the feature selection problem', Optimization Letters, 2(3), pp. 299-308, 2008.

[43]. Agrawal, R., Faloutsos, C., & Swami, A.: 'Efficient similarity search in sequence databases', Springer Berlin Heidelberg, pp. 69-84, 1993.

~~[42].~~[44]. Gelb, A.: 'Applied optimal estimation', MIT press, 1974.

~~[43].~~[45]. Schuldt, C., Laptev, I., Caputo, B.: 'Recognizing human actions: A local SVM approach', in Proc. 17th ICPR, vol. 3, pp. 32–36, 2004.

~~[44].~~[46]. Chen, S., Tian, Y., Liu, Q., Metaxas, DN.: 'Recognizing expressions from face and body gesture by temporal normalized motion and appearance features', Image and Vision Computing. 2013 Feb 28; 31 (2):175-85.

~~[45].~~[47]. Mueid, RM., Ahmed, C., Ahad, MA.: 'Pedestrian activity classification using patterns of motion and histogram of oriented gradient', Journal on Multimodal User Interfaces. pp. 1-7, 2015.