An evaluation of local action descriptors for human action classification in the presence of occlusion

Iveel Jargalsaikhan, Cem Direkoglu, Suzanne Little, and Noel E. O'Connor

INSIGHT Centre for Data Analytics,
Dublin City University, Ireland
iveel.jargalsaikhan2@mail.dcu.ie

Abstract. This paper examines the impact that the choice of local descriptor has on human action classifier performance in the presence of static occlusion. This question is important when applying human action classification to surveillance video that is noisy, crowded, complex and incomplete. In real-world scenarios, it is natural that a human can be occluded by an object while carrying out different actions. However, it is unclear how the performance of the proposed action descriptors are affected by the associated loss of information. In this paper, we evaluate and compare the classification performance of the state-of-art human local action descriptors in the presence of varying degrees of static occlusion. We consider four different local action descriptors: Trajectory (TRAJ), Histogram of Orientation Gradient (HOG), Histogram of Orientation Flow (HOF) and Motion Boundary Histogram (MBH). These descriptors are combined with a standard bag-of-features representation and a Support Vector Machine classifier for action recognition. We investigate the performance of these descriptors and their possible combinations with respect to varying amounts of artificial occlusion in the KTH action dataset. This preliminary investigation shows that MBH in combination with TRAJ has the best performance in the case of partial occlusion while TRAJ in combination with MBH achieves the best results in the presence of heavy occlusion.

1 Introduction

Analyzing complex and dynamic video scenes for the purpose of human action recognition is an important task in computer vision. Therefore, extensive research efforts have been devoted to develop novel approaches for action-based video analysis. Action oriented event detection is an important component for many video management applications especially in surveillance and security [13], sports video [8], and video archive search and indexing domains.

In security applications CCTV footage can be analysed in order to index actions of interest and enable queries relating to actions such as anti-social or criminal behaviour or to monitor crowd volume or agression. This is an especially challenging example of human action recognition due to the volume and quantity

of video and the potentially low level of visual distinctiveness between the actions of interest. This can be seen in the performance of systems used in the TRECVid surveillance event detection (SED) task that has been operating for the last 6 years using the iLIDS dataset from the UK Home Office to annotate video segments with actions such as CellToEar, Embrace, ObjectPut, PeopleMeet, PeopleSplitUp, PersonRuns and Pointing [15]. Some of the unique challenges of this dataset are discussed in [14].

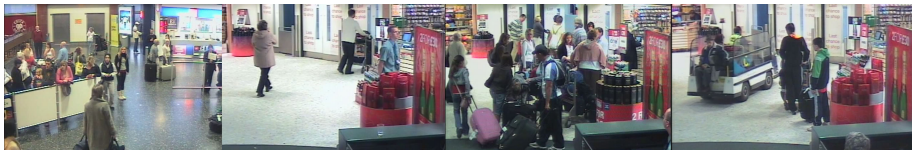


Fig. 1: Sample shots from TRECVid SED dataset show occlusion

Figure 1 shows some example frames from the TRECVid SED dataset illustrating occlusion of the main actor by other objects. Temporal occlusion by other actors (e.g., walking in front of someone who is using a cell phone) is also common. It is difficult to judge the extent of occlusion or the impact of the missing or mis-leading feature descriptors on the performance of human action classifiers trained on example data. Given the size of the TRECVid-SED dataset and the low accuracy levels thus far achieved, we have chosen to use the KTH action dataset to conduct preliminary investigations into the impact of occlusion of human action classification using local descriptors. Although a relatively simple dataset, KTH provides a“ level of playing field ” for testing descriptors.

Despite the fact that existing action description methods have been tested on both artificial and real world datasets, there is no significant study that is directly focused on the problem of occlusion. Occlusion is a challenging problem in real-world scenarios where there are usually many people located at different positions and moving in different individual directions making it difficult to find effective descriptors for higher level analysis.

There are two main classes of human action description methods: global and local. The global methods [4][20] represent the actions based on holistic information about the action and scene. These methods often require the localization of the human body through alignment, background subtraction or tracking. These methods perform well in controlled environments, however exhibit poorer performance in the presence of occlusion, clutter in the background, variance in illumination, and view point changes. Local methods exist [9][12][6] that are less sensitive to these conditions. The local descriptors capture shape and motion information in the neighbourhoods of selected points using image measurements such as spatial or spatio-temporal image gradients and optical flow.

In this paper, we investigate the performance of state-of-art local descriptors for human action recognition in the presence of varying amounts of occlusion. Our objective is to understand how missing action features, i.e. because of static

occlusion, affect action classification performance. In order to model static occlusion, we occlude human action regions with a rectangular shaped, uniform colour object, so that the local descriptors are not extracted within that region.

We evaluate and compare four different local descriptors TRAJ [18], HOG [6], HOF [12], MBH [7] and their possible combinations. These descriptors are combined with a standard bag-of-features representation and a Support Vector Machine (SVM) classifier for action recognition. Our experiments are conducted on the KTH action dataset, and results show that the MBH in combination with TRAJ performs the best in the presence of partial occlusion while TRAJ in combination with MBH achieves the best results in the case of heavy occlusion (greater than 50% of the actor).

To our knowledge, evaluation and comparison of classification performance of local action description methods, in the presence of occlusion, has not been done in the past. However, several authors have evaluated the impact of occlusion on their own work. Weinland et al. [19] showed the robustness of his proposed work under occlusion and view-point changes using artificially imposed occlusions on the KTH and Weizmann datasets. Dollar et al. [10] evaluated the impact of occlusion in terms of pedestrian detection. Additionally, a number of key survey papers in human action recognition [16] [2] [1] stated the necessity of occlusion tolerant action recognition methods. In particular, Poppe [16] wrote "the question [of] how to deal with more severe occlusions has been largely ignored".

The rest of the paper is organized as follows: Section 2 explains the local action descriptors included in our evaluation. Section 3 presents the experimental setup describing how synthetic occlusion is applied to the KTH dataset and evaluation framework. Finally, Section 4 presents and discusses our results prior to the conclusion.

2 Local Action Descriptors

2.1 Trajectory descriptor (TRAJ)

The Trajectory descriptor is proposed in the work of Wang et al. [18]. The descriptor encodes the shape characteristic of a given motion trajectory. Since motion is an important cue in action recognition, this representation allows motion characteristics to be exploited. The descriptor is straight-forward to compute using the points sampled on the trajectory in the image domain. Given a trajectory of length L, the shape is described by a descriptor vector S :

$$S = \frac{\Delta P_t, ..., \Delta P_{t+L-1}}{\sum_{t+L-1}^{j=t} |\Delta P_j|} \tag{1}$$

where $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. In our experiment, the trajectory length was chosen to be $L = 15$ video frames as recommended in [18].

2.2 The HOG/HOF descriptor

The HOG/HOF descriptors were introduced by Laptev et al. in [12]. To characterize local motion and appearance, the authors compute histograms of spatial gradient and optical flow accumulated in space-time neighbourhoods of the selected points. The points can be detected using any interest point detectors [11] [9]. In our experiment, these points are selected along the motion trajectory as in [18]. For the combination of HOG/HOF descriptors with interest point detectors, the descriptor size is defined by $\Delta x(\sigma) = \Delta y(\sigma) = 18\sigma$, $\Delta t(\tau) = 8\tau$. Each volume is subdivided into a $n_x \times n_y \times n_t$ grid of cells; for each cell, 4-bin histograms of gradient orientations (HOG) and 5-bin histograms of optical flow (HOF) are computed. Normalized histograms are concatenated into HOG and HOF as well as HOG/HOF descriptor vectors and are similar in spirit to the well-known SIFT descriptor. In our evaluation we used the grid parameters $n_x = n_y = 3, n_t = 2$ as suggested by the authors [12].

2.3 The Motion Boundary Histogram (MBH) descriptor

Dalal et al. [7] proposed the Motion Boundary Histogram (MBH) descriptor for human detection, where derivatives are computed separately for the horizontal and vertical components of the optical flow. The descriptor encodes the relative motion between pixels. The MBH descriptor separates the optical flow field $I_\omega = (I_x, I_y)$ into its x and y component. Spatial derivatives are computed for each of them and orientation information is quantized into histograms, similarly to the HOG descriptor. We obtain an 8-bin histogram for each component, and normalize them separately with the L_2 norm. Since MBH represents the gradient of the optical flow, constant motion information is suppressed and only information about changes in the flow field (i.e., motion boundaries) is kept. In our evaluation, we used the MBH parameters used in the work of Wang et al. [18].

3 Experimental Setup

3.1 Dataset

The KTH actions dataset [3] consists of six human action classes: walking, jogging, running, boxing, waving, and clapping. Each action class is performed several times by 25 subjects. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The background is homogeneous and static in most of the sequences.

In total, the data consists of 2391 video samples. We follow the original experimental setup of the dataset publishers [3]. Samples are divided into test set (9 subjects: $2, 3, 5, 6, 7, 8, 9, 10$, and 22) and training set (the remaining 16 subjects). We train and evaluate a multi-class classifier and report average accuracy over all classes as the performance measure. The average accuracy is a commonly reported performance measurement when using the KTH dataset [3].

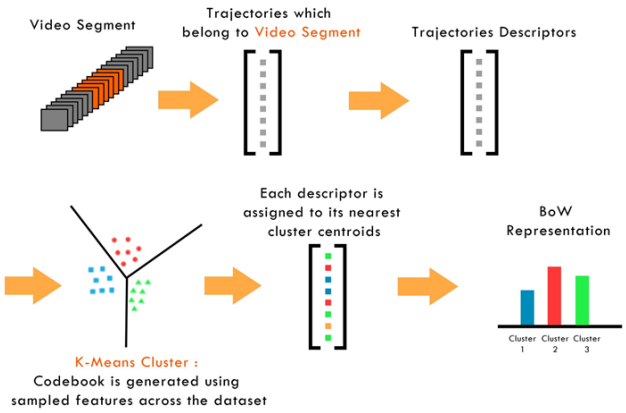


Fig. 2: The extracted trajectory features are represented by each descriptor: TRAJ, HOG, HOF and MBH. Then samples from training videos are used to generate the visual dictionary for respective descriptors. The test video is represented by the Bag-of-Features (BOF) approach using the built visual dictionary. For the case of descriptor combination such as TRAJ+MBH or HOG+HOF, the respective BOF histograms are concatenated together in order to train a SVM classifier.

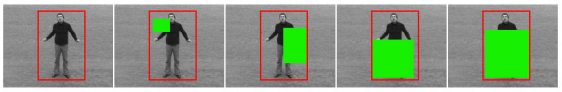


Fig. 3: The sample shots where the different degree of random occlusion is applied into KTH video sequence. The red boundary is manually drawn in order to set an action boundary for each action performer. The green rectangles are occlusion regions randomly selected with 4 different occlusion sizes: 10%, 25%, 50% and 75% of the active region

3.2 Synthetic Occlusion

Occlusion may occur due to static and dynamic occluding objects. For example: If an action performer is occluded by a moving object like a moving car or a person, it is considered as dynamic occlusion. On the other hand, the occluding object may be static like a building or a table then in which case an occlusion represents static occlusion.

In our experiment, we focus our attention on *static occlusion*. Our objective is to understand how the missing action features, i.e. because of *static occlusion*, affect the action classification performance. In order to model static occlusion, we occlude human action regions with rectangular shaped uniform colour objects, so that the action descriptors are not extracted within that regions. The uniform colour ensures no interest points are detected.

Since the KTH action dataset does not contain any occlusion, we have integrated random static occlusion only for the test set sequences. First, action

boundaries are manually selected in each test sequence as a bounding box as shown in red boundary in Figure 3 . The action boundary (AB) should be selected with a specific height H_{AB}, width W_{AB}, position (x_{AB}, y_{AB}), in order to accommodate the region of video where the action is performed. Once we label the action boundaries for all test video sequences, occlusion bounding box (OB) is automatically generated within the action boundary region specified by H_{AB}, W_{AB},x_{AB}, y_{AB} with varying sizes of occlusion area $A(OB)$. The occlusion position is randomly generated and remained static for each test sequence. In our experiment, we have chosen the occlusion areas $A(OB)$ to be 10%, 25%, 50% and 75% of the action boundary area $A(AB)$ as shown in Figure 3. In given action boundary AB and occlusion percentage $Occ\%$, the parameters H_{OB}, W_{OB}, x_{OB}, y_{OB} of the occlusion bounding box OB are randomly selected as follows:

$$\forall H_{OB} \in [H_{AB} - (1 - Occ\%) \times H_{AB}, H_{AB}] \tag{2}$$

$$\forall W_{OB} \in [W_{AB} - (1 - Occ\%) \times W_{AB}, W_{AB}] \tag{3}$$

$$\forall x_{OB} \in [x_{AB}, x_{AB} + (W_{AB} - W_{OB})] \tag{4}$$

$$\forall y_{OB} \in [y_{AB}, y_{AB} + (H_{AB} - H_{OB})] \tag{5}$$

where $Occ\% = \frac{A(OB)}{A(AB)}$ and H_{AB}, W_{AB}, (x_{AB}, y_{AB}) is height, width and top-left corner coordinate of action the boundary box, AB, whereas H_{OB}, W_{OB}, (x_{OB}, y_{OB}) is height, width and top-left corner coordinate of the occlusion boundary box, OB, and $H_{OB}, W_{OB}, x_{OB}, y_{OB} \in \mathbb{N}$.

3.3 Evaluation framework

We adopted the approach of Wang et al. [18] as a video processing pipeline to evaluate spatio-temporal features under different occlusion settings. This approach extracts motion trajectories from the video and generates a set of trajectory with length of $L = 15$ frames.

We compute TRAJ, HOG, HOF and MBH descriptors for each motion trajectory. For volumetric features , HOG, HOF and MBH , we construct 3D volumes along the trajectory. The size of the volume is $N \times N$ pixels and L frames, with $N = 32$ and $L = 15$ in our experiments. The feature vector dimensions of HOG, HOF, MBH and TRAJ are respectively 96, 108, 192 and 30.

In order to represent human actions, we build a Bag-of-Features (BoF) model based on the four different types of descriptors. The Bag-of-Feature representation for each type of descriptor (i.e. HOG, HOF, MBH and TRAJ) is obtained as follows: First, we cluster a subset of $250,000$ descriptors sampled from the training video with the mini batch K-Means algorithm proposed by Sculley [17]. In our experiments, the number of clusters is set to $k = 4,000$, the mini path size is $10,000$ and the number of iterations for clustering is 500. These parameter values are selected empirically to obtain good results and avoid extensive computations. Then each descriptor type is assigned to its nearest cluster centroid using the Euclidean distance to form a co-occurrence histogram.

Descriptor Combination				Recall					Precision				
				No	Partial Occ		Heavy Occ		No	Partial Occ		Heavy Occ	
TRAJ	HOG	HOF	MBH	Occ.	10%	25%	50%	75%	Occ.	10%	25%	50%	75%
			✓	91.2%	89.1%	87.3%	71.8%	49.1%	91.6%	89.8%	88.2%	77.5%	68.3%
		✓		87.0%	87.2%	79.7%	68.5%	45.8%	88.6%	88.0%	81.4%	73.9%	63.6%
		✓	✓	91.2%	89.1%	87.7%	76.9%	50.0%	91.8%	89.8%	88.5%	81.7%	67.6%
	✓			74.5%	69.4%	62.5%	46.8%	26.9%	82.0%	80.8%	74.2%	65.3%	60.4%
	✓		✓	89.8%	88.7%	84.0%	70.4%	46.8%	90.5%	89.6%	85.8%	76.7%	73.9%
	✓	✓		88.4%	87.2%	81.1%	72.2%	44.4%	89.8%	88.7%	83.2%	77.3%	69.1%
	✓	✓	✓	89.8%	89.6%	84.9%	74.1%	49.1%	90.7%	90.6%	86.4%	79.6%	74.4%
✓				87.4%	84.9%	81.5%	79.6%	57.0%	88.7%	86.6%	84.1%	84.3%	73.1%
✓			✓	92.1%	93.4%	86.7%	76.9%	56.5%	92.5%	93.7%	88.1%	82.3%	75.0%
✓		✓		91.2%	88.2%	82.9%	75.5%	52.8%	91.8%	89.0%	84.8%	80.5%	72.0%
✓		✓	✓	92.6%	91.5%	86.2%	76.9%	52.3%	92.9%	92.0%	87.6%	81.2%	70.9%
✓	✓			89.8%	87.8%	81.5%	74.1%	51.9%	90.9%	89.7%	84.9%	80.8%	72.3%
✓	✓		✓	91.6%	90.1%	84.8%	73.6%	51.9%	92.1%	90.9%	86.7%	80.2%	75.9%

Table 1: The precision and recall rate for different combination of our evaluating descriptors. Here, the precision is defined as $P\% = (\frac{TP}{TP+FP}) \times 100$,where TP is true positive, FP is false positive. The Recall (i.e. detection rate) is defined as $R\% = (\frac{TP}{TP+FN}) \times 100$, where TP is true positive and FN is false negative. In this table, all of the measures must be high for a method to show that it can provide sufficient discrimination and classification.

	Descriptor Combination				No	Partial Occlusion		
Rank	TRAJ	HOG	HOF	MBH	Occ.	10%	25%	**Avg.**
1	✓			✓	92.0%	93.4%	86.7%	**90.1%**
2	✓		✓	✓	92.5%	91.5%	86.2%	**88.9%**
3			✓	✓	91.1%	89.1%	87.7%	**88.4%**
4				✓	91.1%	89.0%	87.2%	**88.1%**
5	✓	✓	✓	✓	91.5%	90.5%	84.8%	**87.7%**
6	✓	✓		✓	91.6%	90.1%	84.9%	**87.5%**
7		✓	✓	✓	89.6%	89.5%	84.9%	**87.2%**
8		✓		✓	89.6%	88.5%	84.0%	**86.2%**
9	✓		✓		91.1%	88.4%	83.1%	**85.7%**
10	✓	✓	✓		90.7%	88.8%	82.2%	**85.5%**
11	✓	✓			89.8%	87.8%	81.6%	**84.7%**
12		✓	✓		88.3%	87.3%	81.1%	**84.2%**
13			✓		86.9%	87.3%	79.8%	**83.5%**
14	✓				87.3%	85.0%	81.7%	83.3%
15		✓			74.0%	68.6%	59.4%	64.0%

Table 2: The ranking is computed on the F-Score measure. The F-score is a measure of accuracy that considers precision and recall rates to compute the score as follows: $F\% = 2 \times \frac{Precision \times Recall}{Precision + Recall}$. This table shows the ordered list of descriptor combination in terms their F-Score measure in partial occlusion case. The higher value indicates the better performance

For combining descriptors, we concatenate the co-occurrence histogram of respective descriptors to generate a feature vector to train a SVM classifier. In our evaluation, we train 15 different classifiers for each combination of our four descriptors.

3.4 Classification

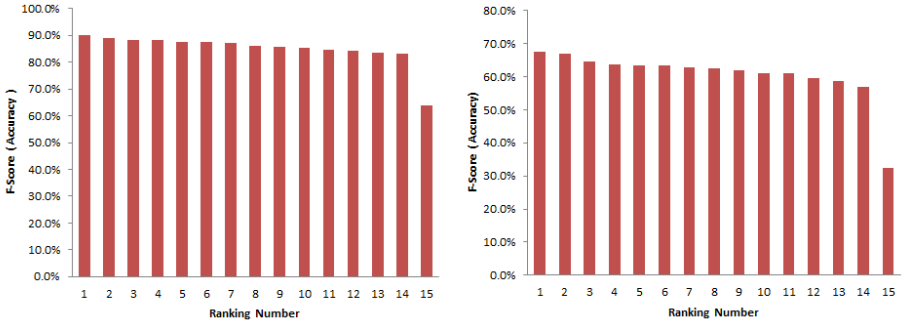
A multi-class support vector machine (SVM) with a Gaussian radial basis function (RBF) kernel is used for classification. We apply a grid searching algorithm to learn the optimal values of the penalty parameter (C) in SVM and the scaling factor (γ) in Gaussian RBF kernel with the KTH dataset training set (without any occlusion). The grid searching is performed using 10 fold cross-validation. The optimal parameter values are : $C = 1$ and $\gamma = 32 \times 10^{-2}$. These parameters are fixed throughout our evaluation of the local descriptors and their possible combinations.

	Descriptor Combination				No	Heavy Occlusion		
Rank	TRAJ	HOG	HOF	MBH	Occ	50%	75%	**Avg.**
1	✓				87.3%	79.2%	56.1%	**67.7%**
2	✓			✓	92.0%	76.7%	57.2%	**66.9%**
3	✓		✓	✓	92.5%	76.6%	52.7%	**64.7%**
4	✓		✓		91.1%	74.9%	52.8%	**63.8%**
5	✓	✓		✓	91.6%	73.7%	53.0%	**63.3%**
6			✓	✓	91.1%	77.0%	49.7%	**63.3%**
7	✓	✓			89.8%	74.0%	51.5%	**62.8%**
8		✓	✓	✓	89.6%	74.3%	50.8%	**62.5%**
9	✓	✓	✓	✓	91.6%	73.8%	50.2%	**62.0%**
10				✓	91.1%	72.0%	50.1%	**61.1%**
11	✓	✓	✓		90.7%	74.3%	47.8%	**61.1%**
12		✓		✓	89.6%	70.8%	48.5%	**59.7%**
13		✓	✓		88.3%	72.3%	45.3%	**58.8%**
14			✓		86.9%	68.3%	45.8%	**57.0%**
15		✓			74.0%	42.2%	22.5%	**32.3%**

Table 3: Here shows the F-Score based ranking in heavy occlusion case for local action descriptors and their possible combinations.

4 Experimental Results

Table 2 shows the ranking of different combinations of descriptors in the partial occlusion case based on F-Score. The best three combinations are TRAJ+MBH (90.1 %), TRAJ+HOF+MBH (88.9%) and HOF+MBH (88.4%). The worst performance is with HOG and HOF features. HOG descriptor obtained 64.3% and HOF descriptor obtained 83.3% and their combination is 83.5%.



(a) Partial Occlusion (b) Heavy Occlusion

Fig. 4: The graphical illustration of accuracy for partial and heavy occlusion cases. (a) The partial occlusion case. The ranking number corresponds to Table 2 (b) The heavy occlusion case. The ranking number corresponds to Table 3

The heavy occlusion ranking is presented in Table 3. TRAJ (67.7%), TRAJ+MBH (66.9%), TRAJ+HOF+MBH (64.7%) combinations perform best. The HOG, HOF and their combination perform poorly. Generally, the best descriptors are TRAJ, MBH and their combination. They consistently outperform any other combination for different scales of occlusion area in our experiments.

We now present experimental results for various descriptor combinations. We use multi-class classification where we apply the one-against-rest approach and compare the performance based on precision, recall and F-score. The scores are reported as an average of the 6 action classes. In order to measure the occlusion impact, we compute the above mentioned scores at four different cases of occlusion: 10%, 25%, 50% and 75% occlusion of the action area. We also group the cases into partial occlusion (10% 25% occluded) and heavy occlusion (50% 75% occluded). The classifier is trained with non-occluded training data. Therefore all occlusion cases are classified with the same trained classifier.

Table 1 shows the recall and precision scores for all combinations of the descriptors we evaluated. The recall is calculated for partial and heavy occlusion scenarios. In partial occlusion, MBH and its combination with other descriptors performed significantly better than other combinations. Especially the combination of TRAJ + MBH outperforms the without-occlusion case by 2%. This can be explained by the fact that occlusion also acts like a noise filtering. It increases the discriminative power of the representation. Regarding the heavy occlusion, the best performance is shown with all four combinations of trajectory descriptor. It makes the trajectory descriptor particularly suitable for scenarios with large occlusions. For example, with 75% occluded area, TRAJ individually obtained 57% recall rate which is the highest score compared to any other combination where most of them barely reached 50%.

In terms of precision, the same trend is observed in both occlusion scenarios. The partial occlusion is predominantly handled significantly better than others

when there is combination of MBH descriptors. For heavy occlusion, TRAJ + MBH descriptors topped the precision rank.

The poorest performance is exhibited by HOG and its combination with other descriptors. In both partial and heavy occlusion cases, the HOG descriptor obtained the worst precision and recall rate. Therefore it is unsuitable to use HOG even with other occlusion tolerant features like MBH or TRAJ as it significantly decreases the performance.

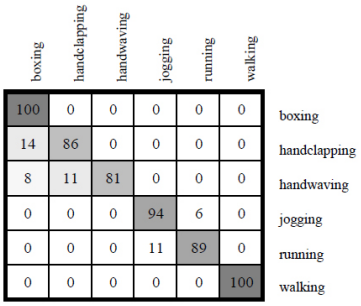


Fig. 5: Confusion matrix for the un-occluded KTH dataset

5 Discussion

The experimental results confirm that the motion based descriptors (TRAJ, HOF and MBH) are more discriminative when recognizing human actions in an occluded scene. Among the motion based descriptors, MBH and TRAJ descriptors significantly outperform other descriptors. In the partial occlusion case, MBH is the best choice, whereas the TRAJ descriptor is good for heavy occlusion. Texture or appearance based descriptors (HOG) performed poorly in the presence of occlusion because the objects shape undergoes significant changes due to the occlusion artefact. We observed that combining MBH and TRAJ descriptors outperforms other possible combinations in both partial and heavy occlusion.

The performance under very heavy occlusion in particular is surprising. While showing a significant decrease in performance compared with no occlusion, average precision over the six actions of greater than 60% is still achieved. We speculate that this is due to the extremely simplified nature of the KTH dataset, a facet noted in a recent review of datasets for human action recognition [5] that described the unrealistic nature of KTH. The differentiation between classes is high (see the confusion matrix for the baseline classification with no occlusion, Figure 5) and the area of the action boundary is relatively large. Therefore actions can be successfully differentiated by the multi-class classifier using only a small number of local descriptors.

Performance with heavy occlusion in real-world surveillance datasets is predicted to be very poor. However the strong performance of the MBH descriptor either alone or combined with TRAJ is likely to transfer to the more complex scenes.

6 Conclusion and Future Work

We have presented an evaluation and comparison framework for the state-of-art human local action descriptors. We evaluated four different local action descriptors which are Trajectory (TRAJ), Histogram of Orientation Gradient (HOG),Histogram of Orientation Flow (HOF) and Motion Boundary Histogram (MBH). These descriptors are experimented with a standard bag-of-features representation and a Support Vector Machine classifier. We investigate the performance of these descriptors and their possible combinations with respect to varying amount of artificial occlusion in the KTH action dataset. Results show that the MBH and its combination with TRAJ achieve the best performance in partial occlusion. TRAJ and its combination with MBH perform the best results in the presence of heavy occlusion.

Indictations regarding the relative importance and robustness of local action descriptors will assist in designing systems that are more resilliant to the frequent occurences of occlusion. Particularly in developing classifiers for the more complex actions and scenes found in surveillance and security applications. We hope that weighting local action descriptors in scenarios where higher levels of occlusion are likely (such as the scenes shown in figure 1) will improve the overall accuracy of the classifier.

This work demonstrated that the choice of local descriptor has an impact on the classifier performance in the presence of occlusion. Further work will explore how this will transfer to real-world applications. Particularly we will expand our evaluation with more descriptors, as well as real-world datasets like TRECVid-SED [15] and Hollywood [11] with examples of realistic occlusion.

Acknowledgements

References

1. J. K. Aggarwal and Q. Cai. Human motion analysis: A review. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 90–102. IEEE, 1997.
2. L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, 2011.
3. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE, 2005.

4. A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001.
5. J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 2013.
6. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.
7. N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *Computer Vision–ECCV 2006*, pages 428–441, 2006.
8. C. Direkoglu and N. O'Connor. Team activity recognition in sports. *ECCV 2012*, pages 69–83, 2012.
9. P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.
10. P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009.
11. I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2):107–123, 2005.
12. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Computer Society Conference on*, pages 1–8, 2008.
13. M.-Y. Liao, D.-Y. Chen, C.-W. Sua, and H.-R. Tyan. Real-time event detection and its application to surveillance systems. In *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pages 4–pp. IEEE, 2006.
14. S. Little, I. Jargalsaikhan, K. Clawson, M. Nieto, H. Li, C. Direkoglu, N. E. O'Connor, A. F. Smeaton, B. Scotney, H. Wang, and J.Liu. An information retrieval approach to identifying infrequent events in surveillance video. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 223–230. ACM, 2013.
15. P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, G. Quénot, et al. An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2011-TREC Video Retrieval Evaluation Online*, 2011.
16. R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.
17. D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178. ACM, 2010.
18. H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *IEEE CVPR*, pages 3169–3176, 2011.
19. D. Weinland, M. Özuysal, and P. Fua. Making action recognition robust to occlusions and viewpoint changes. *Computer Vision–ECCV 2010*, pages 635–648, 2010.
20. A. Yilmaz and M. Shah. A differential geometric approach to representing the human actions. *Computer Vision and Image Understanding*, 109(3):335–351, 2008.