Recognising Action as Clouds of Space-Time Interest Points

Matteo Bregonzio, Shaogang Gong and Tao Xiang
School of Electronic Engineering and Computer Science
Queen Mary University of London, London E1 4NS, United Kingdom
{bregonzio,sgg,txiang}@dcs.qmul.ac.uk

Abstract

Much of recent action recognition research is based on space-time interest points extracted from video using a Bag of Words (BOW) representation. It mainly relies on the discriminative power of individual local space-time descriptors, whilst ignoring potentially valuable information about the global spatio-temporal distribution of interest points. In this paper, we propose a novel action recognition approach which differs significantly from previous interest points based approaches in that only the global spatio-temporal distribution of the interest points are exploited. This is achieved through extracting holistic features from clouds of interest points accumulated over multiple temporal scales followed by automatic feature selection. Our approach avoids the non-trivial problems of selecting the optimal space-time descriptor, clustering algorithm for constructing a codebook, and selecting codebook size faced by previous interest points based methods. Our model is able to capture smooth motions, robust to view changes and occlusions at a low computation cost. Experiments using the KTH and WEIZMANN datasets demonstrate that our approach outperforms most existing methods.

1. Introduction

Human action recognition is one of the most active research areas in computer vision with many real-world applications, such as human-computer interaction, video indexing, video surveillance and sport events analysis. It is a challenging problem as actions can be performed by subjects of different size, appearance and pose. The problem is compounded by the inevitable occlusion, illumination change, shadow, and camera movement.

Early work on action recognition is based on tracking [16, 1, 15, 19] or spatio-temporal shape template [7, 8, 21]. Both tracking and spatio-temporal shape template construction require highly detailed silhouettes to be extracted, which may not be possible given a real-world noisy video input. To address this problem, space-time interest point based approaches have become increasingly popular [3, 18, 11, 17]. These approaches are based on a Bag of Words (BOW) feature representation that has been successfully applied to 2-D object categorisation and recognition. Compared to tracking and spatio-temporal shape based approaches, they are more robust to noise, camera movement, and low resolution inputs. Nevertheless, they rely solely on the discriminative power of individual local space-time descriptors. Information about the global spatio-temporal distribution of the interest points is ignored. Consequently, they are unable to capture smooth motions due to lack of temporal information. Furthermore, their have to address the non-trivial problems of selecting the optimal space-time descriptor, clustering algorithm for constructing a codebook and codebook size, which inevitably involve parameter tuning. Such parameter settings are highly data dependent and re-tuning is required for different video inputs.

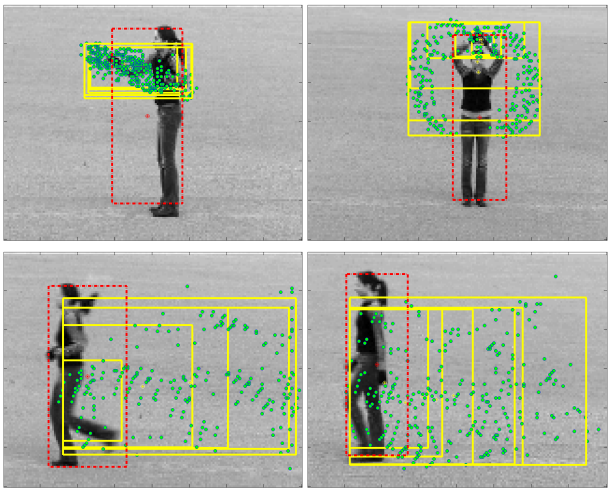


Figure 1. Examples of clouds of interest points. The clouds at different temporal scales are highlighted in yellow boxes.

In this paper, we propose a novel approach based on representing action as clouds of interest points accumulated at different temporal scales. Specifically, a new space-time interest point detection method is developed to extract denser

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

and more informative interest points compared to the existing interest point extraction methods [3, 18]. In particular, our model avoids spurious detection in both background areas and highly textured static foreground areas unrepresentative of the dynamic parts of actions concerned. The extracted interest points are accumulated over time at different temporal scales to form point clouds. Examples of the clouds of interest points of different temporal scales are shown in Fig. 1. Holistic features are then computed from these point clouds for action representation, which capture *explicitly* and *globally* the spatial and temporal distribution of salient local space-time patches. A simple yet effective feature selection method is then formulated to select the most informative features for discriminating different classes of actions.

Our approach differs significantly from the conventional Bag of Words (BOW) based approaches in that a completely different aspect of interest points from video is exploited. More specifically, the conventional BOW methods are focused on what these interest points are and how often they occur in the video (i.e. they rely primarily on the discriminative power of individual local space-time descriptors). Information about the spatial and temporal distribution of interest points is lost. In contrast, our approach exploits only the global spatio-temporal information about where the interest points are and when they are detected. Without the need to represent the detected interest points using local descriptors and classify them into video-words, ad hoc and arbitrary parameter tuning process is avoided. In addition, our model is novel in capturing information about global spatio-temporal distribution of interest points explicitly and at different scales. It therefore is able to capture smooth motions, robust to view changes and occlusions, and with a low computation cost. The proposed approach is evaluated using the KTH dataset [18] and the WEIZMANN dataset [7]. Experimental results demonstrate that our model outperforms most existing techniques.

2. Related Work

Existing human action recognition methods can be broadly classified into four categories: flow based [4], spatio-temporal shape template based [7, 8, 21], interest points based [3, 18, 11, 17, 10, 6, 22], and tracking based [16, 1, 15, 19]. Flow based approaches construct action templates based on optical flow computation [4, 5]. However, the features extracted from the flow templates are sensitive to noise especially at the boundary of the segmented human body. Spatio-temporal shape template based approaches essentially treat the action recognition problem as a 3-D object recognition problem by representing action using features extracted from spatio-temporal volume of an action sequence [7, 8, 21]. These techniques require highly detailed silhouettes to be extracted, which may not be possible given a real-world noisy video input. In addition, the computational cost of space-time volume based approaches is unacceptable for real-time applications. Tracking based approaches [16, 1, 15, 19] suffer from the same problems. Consequently, although 100% recognition rate has been reported on the 'clean' WEIZMANN dataset, these approaches mostly fail on a noisy dataset such as the KTH dataset, which is featured with low resolution, strong shadows, and camera movement rendering clean silhouette extraction impossible.

To address this problem, Schüldt et al. [18] propose to represent action using 3-D space-time interest points detected from video. The detected points are clustered to form a dictionary of prototypes or video-words. Each action sequence is then represented by Bag of Words. Dollar et al. [3] introduce a multidimensional linear filter detector, which results in the detection of denser interest points. However, their methods ignore information about the global spatio-temporal distribution of the interest points. Consequently, they are unable to capture smooth motions due to lack of temporal information. This also explains why they generate poor results on the clean yet more ambiguous WEIZMANN dataset whilst working reasonably well on the KTH dataset.

To overcome the limitations of the conventional BOW model, a number of recent attempts have been made to utilise information about the spatio and temporal distribution of interest points. Liu and Shah exploit the spatial distribution of interest points using a modified correlogram [10]. Gilbert et al. [6] encode spatial information through concatenating video-words detected at different regions. Zhang et al. [22] introduce the concept of motion context to capture both spatial and temporal distribution of video-words.

All these extensions, however, still suffer from some of the inherent flaws of the original BOW method, in that ad hoc and arbitrary processes are needed for selecting data dependent optimal space-time descriptor, clustering algorithm for constructing a codebook, and codebook size. In addition, spatial and temporal information about the distribution of interest points are only exploited implicitly, locally, and at a fixed temporal scale. In contrast, our model avoids data specific parameter tuning and exploits spatio-temporal information explicitly and at multiple temporal scales therefore capturing both local and global temporal information about interest points distribution.

3. Interest points Detection

Interest points are local spatio-temporal features considered to be salient or descriptive of the action captured in a video. Among various interest point detection methods, the one proposed by Dollar et al. [3] is perhaps the most widely used for action recognition. Using their detector, intensity

(a) Boxing (b) Hand waving (c) Running

Figure 2. Comparison between interest points detected using our detector (green circle points) and the Dollar et al. [3] detector (red square points).

variations in the temporal domain are detected using Gabor filtering. The detected interest points correspond to local 3-D patches that undergo complex motions. Specifically, the response function of the Gabor filters is given as:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \quad (1)$$

where $g(x, y : \sigma)$ is the Gaussian smoothing kernel to be applied in the spatial domain, and h_{ev} and h_{or} are the 1D Gabor filters applied temporally, defined as:

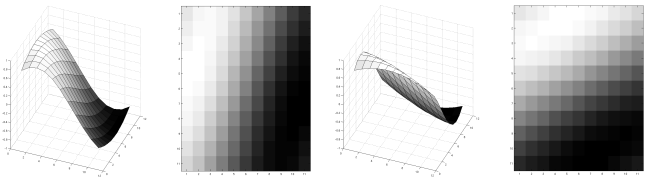
$$h_{ev}(t; \tau, \omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2} \quad (2)$$

$$h_{od}(t; \tau, \omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2} \quad (3)$$

By setting $\omega = 4/\tau$, there are essentially two free parameters τ and σ which roughly control the spatial and temporal scales of the detector.

Despite its popularity, the Dollar detector has a number of drawbacks: it ignores pure translational motions; since it uses solely local information within a small region, it is prone to false detection due to video noise; it also tends to generate spurious detection background area surrounding object boundary and highly textured foreground areas; it is particularly ineffective given slow object movement, small camera movement, or camera zooming. Some of the drawbacks are highlighted in the examples shown in Fig. 2.

To overcome these shortcomings, we propose here a different interest point detector. The shortcomings of the Dollar detector are caused mostly by its design of spatial and temporal filters and the way these filters are combined to give the final response. In particular, the 1-D Gabor filter



(a) (b)

Figure 3. Examples of the 2D Gabor filters oriented along (a) 22° and (b) 67°.

applied in the temporal domain is sensitive to both background noise and highly textured object foreground areas regardless their relevance to capturing the dynamics of actions observed. To overcome this problem, the proposed detector explores different filters for detecting salient space-time local areas undergoing complex motions. More specifically, our detector consists of two steps: 1) frame differencing for focus of attention and region of interest detection, and 2) Gabor filtering on the detected regions of interest using 2D Gabor filters of different orientations. This two-steps approach facilitates saliency detection in both the temporal and spatial domains to give a combined filter response. The 2D Gabor filters are composed of two parts. The first part $s(x, y; i)$ represents the real part of a complex sinusoid, known as the carrier:

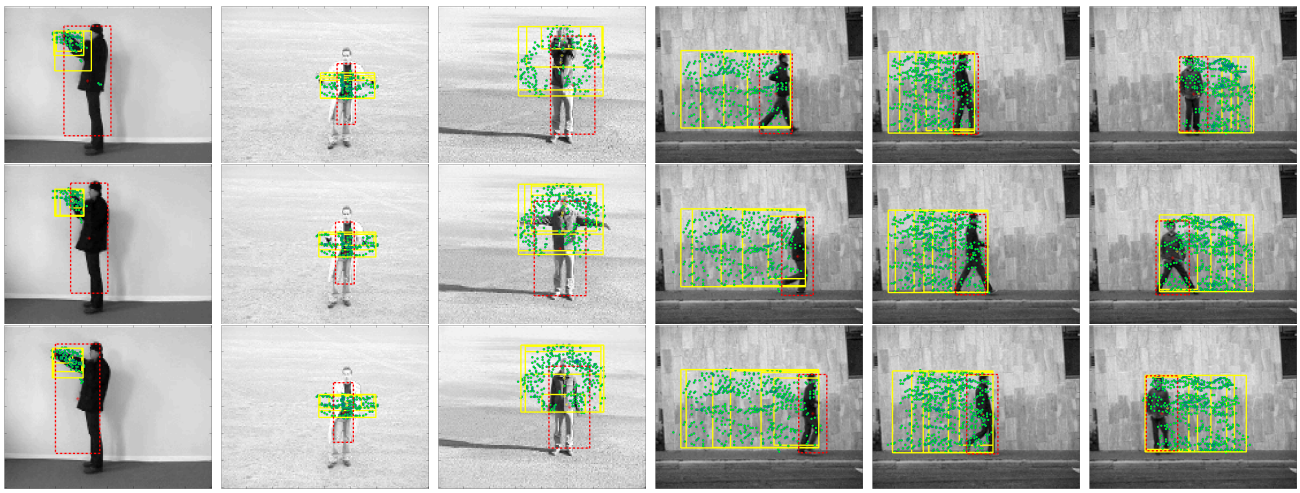
$$s(x, y; i) = cos(2\pi(\mu_0 x + \upsilon_0 y) + \theta_i) \quad (4)$$

where θ_i defines the orientation of the filter and 5 orientations are considered: $\theta_{i=1,..,5} = \{0^\circ, 22^\circ, 45^\circ, 67^\circ, 90^\circ\}$, and μ_0 and υ_0 are the spatial frequencies of the sinusoid controlling the scale of the filter. The second part of the filter $G(x, y)$ represents a 2D Gaussian-shaped function, known as the envelope:

$$G(x, y) = exp\left(-\frac{\frac{x^2}{\rho^2} + \frac{y^2}{\rho^2}}{2}\right) \quad (5)$$

where ρ is the parameter that controls the width of $G(x, y)$. We have $\mu_0 = \upsilon_0 = \frac{1}{2\rho}$; therefore the only parameter controlling the scale is ρ, which is set to 11 pixels in this study.

Fig. 2 shows examples of our interest point detection results using the KTH dataset. It is evident that the detected interest points are much more meaningful and descriptive compared to those detected using the Dollar detector. In particular, the interest points detected by our approach tend to correspond to the main contributing body parts to the action being performed whilst those detected by Dollar detect often drift to static body parts of high texture or background with strong edges.



(a) Boxing (b) Hand clapping (c) Hand waving (e) Running (f) Walking (g) Gallop sideways

Figure 4. Examples of clouds of space-time interest points. We have $S = 6$ and $N_s = 5$. In each frame the red rectangle represents the foreground area, the green points are the extracted interest points, and the yellow rectangles illustrate clouds of different scales.

4. Action Representation

4.1. Clouds of Interest Points

Suppose an action video sequence $\mathbf{V}$ consisting of T image frames, represented as $\mathbf{V} = [\mathbf{I}_1, \ldots, \mathbf{I}_t, \ldots, \mathbf{I}_T]$, where $\mathbf{I}_t$ is the tth image frame. For the image frame $\mathbf{I}_t$, a total of S interest point clouds of different temporal scales are formed. They are denoted as $[\mathbf{C}_t^1, \ldots, \mathbf{C}_t^s, \ldots, \mathbf{C}_t^S]$. More specifically, the cloud of the s-th scale is constructed by accumulating the interest points detected over the past $s \times N_s$ frames, where N_s is the difference between two consecutive scales (in the number of frames). Examples of the clouds of interest points extracted from the KTH and WEIZMANN datasets are shown in Fig. 4. It is evident that different types of actions exhibit interest point clouds of very different shape, relative location, and distribution. It is also evident that interest point clouds of different scales capture different aspects of the human motion which potentially have different levels of discriminative power. This will be addressed by feature selection detailed in Sec. 4.3.

4.2. Feature Extraction

For each image $\mathbf{I}_t$, two sets of features are extracted. These features are significantly different from the local descriptors computed by the conventional interest point based approaches. The former are global and holistic, while the latter, computed from a cuboid centred at each interest point, are local.

The first set is concerned with the shape and speed of the foreground object. To reliably detect and segment a foreground object given camera movement and zooming, strong shadow, and noisy input is a non-trivial task. This is accomplished by the following procedure. Firstly, regions of interest are detected via frame differencing. Secondly, a series of 2-D Gabor filters are applied to the image frame. Thirdly, the responses of these filters are fused together with the frame differencing result. Finally, a Prewitt edge detector [13] is employed to segment the object from the detected foreground area. Once the object is segmented from the frame, two features are computed: O_t^r measuring the height and width ratio of the object, and O_t^{Sp} measuring the absolute speed of the object.

The second set of features are extracted from the interest point clouds of different scales. These features are thus scale dependent. Particularly, from the s-th scale cloud, 8 features are computed and denoted as

$$[C_s^r, C_s^{Sp}, C_s^D, C_s^{Vd}, C_s^{Hd}, C_s^{Hr}, C_s^{Wr}, C_s^{Or}] \qquad (6)$$

Note that subscript t is omitted for clarity. Specifically, C_s^r is the height and width ratio of the cloud. C_s^{Sp} is the absolute speed of the cloud. C_s^D is the density of the interest point within the cloud, which is computed as the total number of points normalised by the area of the cloud. C_s^{Vd} and C_s^{Hd} measure the spatial relationship between the cloud and the detected object area. Specifically, C_s^{Vd} is the vertical distance between the geometrical centre (centroid) of the object area and the cloud, and C_s^{Hd} is the distance in the horizontal direction. C_s^{Hr} and C_s^{Wr} are the height ratio and width ratio between the object area and the cloud respectively. The amount of overlap between the two areas is measured by C_s^{Or}. Overall, the 8 features can be put into two categories: C_s^r, C_s^{Sp}, and C_s^D measure the shape, speed and density of cloud itself; the rest 5 features capture the relative shape and location information between the object and the cloud areas. To make these features insensitive

to outliers in the detected interest points, an outlier filter is deployed before the feature extraction, which evaluates the interest point distribution over 4 consecutive frames and removes those points that are too far away from the distribution boundaries.

Now each frame is represented using $8S + 2$ features where S is the total number of scales (i.e. 8 features for each scale plus 2 scale-independent features O_t^r and O_t^{Sp}). To represent the whole action sequence, a total of $(8S+2) \times T$ features need be used which leads to a feature space of a very high dimension that can cause overfitting giving poor recognition performance. To reduce the dimensionality of the feature space, and more importantly, to make our representation less sensitive to feature noise and invariant to the duration T of each action sequence, a histogram of N_b bins is constructed for each of the $8S + 2$ features collected over time via linear quantization. Consequently, each action sequence is represented as $8S+2$ histograms or $(8S+2) \times N_b$ features whilst $N_b \ll T$. This reduced feature space dimension is still very high. Moreover, there are uninformative and redundant features one would wish to eliminate from the feature set. To that end, a simple and intuitive yet effective feature selection method is formulated as follows.

4.3. Feature Selection

We introduce a feature selection method that measures the relevance of each feature according to how much its value varies within each action class and across different classes. Specifically, a feature is deemed as being informative and relevant to the recognition task if its value varies little for actions of the same class but varies significantly for actions of different classes. First, given a training set of K action classes, for the i-th feature f_i in the k-th class, we compute its mean and standard deviation within the class as $\mu_{f_i}^k$ and $\sigma_{f_i}^k$ respectively. The relevant measure for feature f_i is then denoted as R_{f_i} and computed as:

$$R_{f_i} = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mu_{f_i}^k - \frac{1}{K}\sum_{k=1}^{K}\mu_{f_i}^k)^2}}{\frac{1}{K}\sum_{k=1}^{K}\sigma_{f_i}^k} \qquad (7)$$

The numerator and denominator of the above equation correspond to the standard deviation of the intra class mean, and the inter class mean of the intra class standard deviations respectively. The former measures how the feature value varies across different classes (the higher the value is, the more informative the feature f_i is); the latter tells us how the value varies within each class (the lower the value, the more informative the feature). Overall features with higher R_{f_i} values are preferred over those with lower ones. Finally, all features are ranked according to their R_{f_i} and a decision is made as to what percentage of the features are to be kept for recognition.

Even though this feature selection method is relevance based, it does not consider explicitly mutual information, such as in [14]. In our model, different features are selected separately as if they are independent of each other. It has been shown that combining good features does not guarantee good recognition performance [14]. This suggests that one would want to select features collectively. However, in our case the feature search space is potentially very high for exhaustive search. Even a sequential-search based approximation scheme can be overly expensive. A significant advantage of our method is its extremely low computational cost. We show empirically through experiments that our method is more effective than a far more sophisticated method using mutual information [14].

5. Experiments

5.1. Datasets

KTH Dataset – The KTH dataset was provided by Schuldt et al. [18] in 2004 and is the largest public human activity video dataset. It contains 6 types of actions (boxing, hand clapping, hand waving, jogging, running and walking) performed by 25 subjects in 4 different scenarios including indoor, outdoor, changes in clothing and variations in scale. Each video clip contains one subject performing a single action. Each subject is captured in a total of 23 or 24 clips, giving a total of 599 video clips. Each clip is sampled at 25Hz and lasts between 10 to 15 seconds with image frame size of 160×120 (see examples in Fig. 5).

WEIZMANN Dataset – The WEIZMANN dataset was provided by Blank et al. [2] in 2005. It contains 90 video clips from 9 different subjects. Again, each video clip contains one subject performing a single action. There are 10 different action categories: walking, running, jumping, gallop sideways, bending, one-hand-waving, two-hands-waving, jumping in place, jumping jack, and skipping. Each clip lasts about 2 seconds at 25Hz with image frame size of 180×144 (see examples in Fig. 5).

A robustness test dataset is also provided by the same WEIZMANN group. It consists of 11 walking sequences with partial occlusions and non-rigid deformations (e.g. walking in skirt, walking with a briefcase, knees up walking, limping man, occluded legs, walking swinging a bag, sleepwalking, walking with a dog). The dataset also includes 9 walking sequences captured from different viewpoints (from 0° to 81° with 9° increments from the horizontal plane). This dataset is designed to test model robustness under occlusion, view variation and non-rigid deformation (see examples in Fig. 5).

5.2. Recognition Settings

Recognition was performed using Nearest Neighbour Classifier (NNC) and Support Vector Machine (SVM), also used widely elsewhere for action recognition. For NNC, absolute distance was used. For SVM, the radial basis function kernel was used.

Figure 5. From top to bottom: example frames from KTH dataset, WEIZMANN dataset and WEIZMANN robustness test dataset.

Our approach was validated using Leave-One-Out Cross-Validation (LOOCV). It involved employing a group of clips from a single subject in a dataset as the testing data, and the remaining clips as the training data. This was repeated so that each group of clips in the dataset is used once as the testing data. More specifically, for the KTH dataset the clips of 24 subjects were used for training and the clips of the remaining subject were used for validation. For the WEIZMANN dataset, the training set contains 8 subjects. As for the WEIZMANN robustness test dataset, the whole WEIZMANN action recognition dataset was used as training set. Each of the 20 robustness test sequences was classified as one of the 10 action classes.

For constructing the multi-scale interest point clouds, N_s was set to 5 and the total number of scales was 6. This gives to 50 features each represented as a 90-bin histogram through linear quantisation (i.e. the features are represented in a 4500 dimensional space). Our feature selection model removed 20% of these features.

Note that the existing Bag of Words methods require generating a codebook using a clustering algorithm such as k-means which is sensitive to initialisation. Therefore typically results were reported in the literature based on average of 20 trials. In our method no such initialisation issue exists, so different trials will give an identical result.

5.3. Recognition Rate

Our experimental results show that NNC and SVM give similar performance using our features, with NNC slightly outperforming SVM. The results are shown in Fig. 6. In particular, we obtained a recognition rate of 93.17% for KTH dataset and 96.66 % for WEIZMANN dataset. Table 1 also compares our results with the existing approaches proposed recently, which are not restricted to interest points based methods. It shows that our results are close to the best result reported so far on each dataset, and outperforms most recently proposed methods, especially those tested on both datasets.

METHOD	KTH	WEZIMANN
Our approach	**93.17%**	**96.66%**
Fathi et al. [5]	90.5%	100%
Zhang et al. [22]	91.33%	92.89%
Kläser et al. [9]	91.4%	84.3%
Niebles et al. [11]	83.3%	90.0%
Dollar et al. [3]	81.17%	85.2%
Liu et al. [10]	94.16%	-
Zhao et al. [23]	91.17%	-
Gilbert et al. [6]	89.92%	-
Savarese et al. [17]	86.83%	-
Nowozin et al.[12]	84.72%	-

Table 1. Comparative results on the KTH and WEIZMANN datasets.

5.4. Robustness Test

We demonstrate the robustness of our method using the WEIZMANN robustness test sequences. Examples of the detected clouds of interest points are shown in Fig. 7. Only 1 of the 20 sequences was wrongly classified. This sequence contains a person walking with a dog and was recognised as skipping. The most informative human body part for the action (i.e. the legs) overlapped with another object (the dog), which was also walking but in a very different way (see Fig. 7(d)). This sequence is therefore challenging for any existing recognition methods. In comparison, the method in [2] also obtained the same result as us with the same sequence miss-classified. Our method outperforms the method in [20] which wrongly classified 2 sequences as well as Dollar's method [3] which, over 20 runs, classifies less than one of the 20 sequences correctly on average. To the best of our knowledge, no other action recognition approach has reported result on this robustness test dataset.

Among the three methods we compared for robustness,

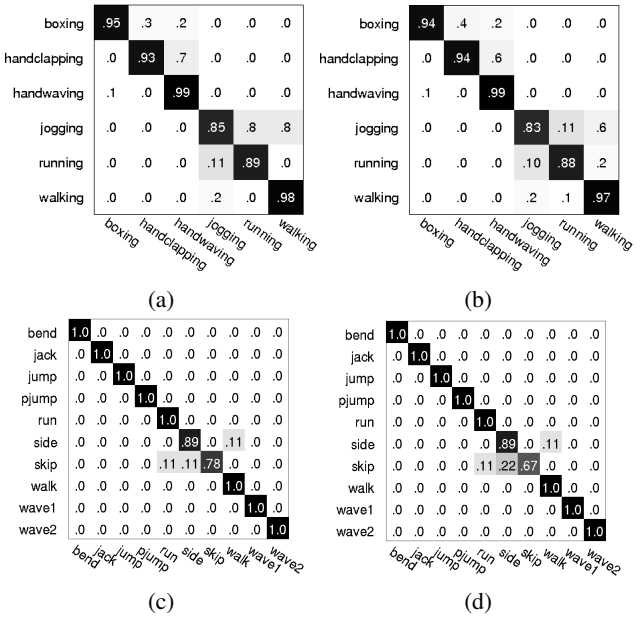
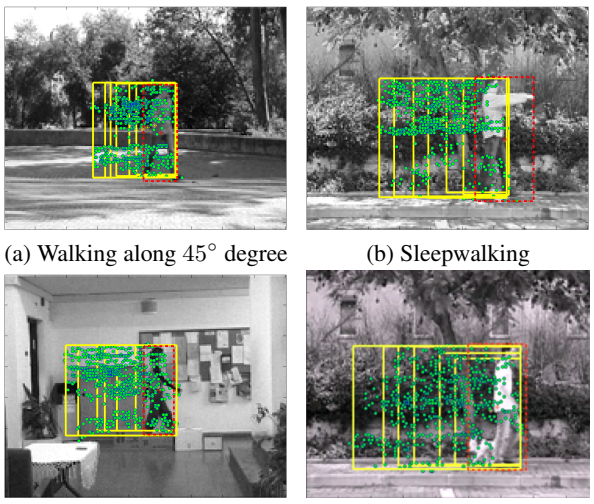


Figure 6. Recognition performance of our approach measured using confusion matrices: (a) KTH dataset using NNC recognition algorithm, accuracy: 93.17% (b) KTH dataset using SVM recognition algorithm, accuracy: 92.5% (c) WEIZMANN dataset using NNC recognition algorithm, accuracy: 96.66% (d) WEIZMANN dataset using SVM recognition algorithm, accuracy: 95.55%

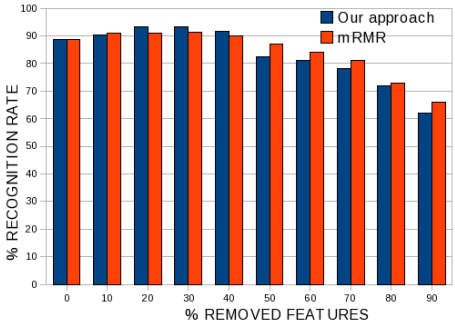


(a) Walking along 45° degree (b) Sleepwalking

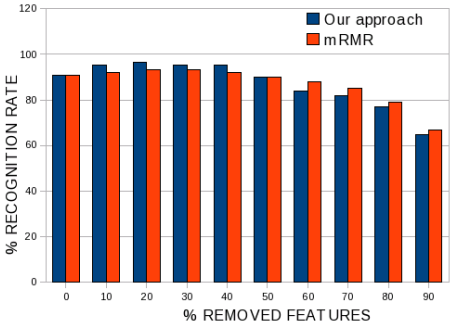
(c) Walking with occluded legs (d) Walking with a dog

Figure 7. Example of clouds of points detected in the sequences used in the robustness test experiments.

the methods in [2] and [20] are based on space-time volume representation, whilst the one in [3] is based on Bags of Words (BOW) representation using interest points. Our experiments show that by modelling explicitly global spatial and temporal distribution of interest points, our approach is significantly more robust than a conventional BOW based



(a) KTH dataset



(b) WEIZMANN dataset

Figure 8. Comparing our feature selection method with the mRMR method in [14]. The two methods were compared with different percentages of less-informative features were eliminated (from 0% to 90 %).

method and is comparable to space-time volume based methods. However, the space-time volume based methods [2, 20] are more sensitive to input noise as they rely on accurate silhouette extraction. They are thus expected to perform poorly on the more noisy KTH dataset.

5.5. Effect of Feature Selection

Fig. 8 shows how the recognition performance of our approach is affected when different amount of features are removed using our feature selection method. Our method was also compared with a more complex minimal-redundancy-maximal-relevance (mRMR) algorithm proposed in [14]. Fig. 8 shows that our method outperforms the mRMR method. A major attraction of the mRMR method, as compared to other existing feature selection methods, is its low computational cost. In comparison, our method takes less than one tenth of the time used by the mRMR method for selecting the same amount of features. It took 9.2 seconds using our method for measuring and ranking 4500 features, as compared to 131.5 seconds using mRMR on the same PC Workstation setup.

To evaluate how much the feature selection step has contributed to the final recognition performance, we carried out experiment to evaluate our approach without feature selec-

METHOD	FEATURE SELEC.	KTH	WEIZMANN
Our approach	NO	90.1%	93.2%
	YES	93.17%	96.6%
Dollar et al. [3]	NO	81.17%	85.2%
	YES	86.6%	87.3%

Table 2. Evaluation of the effectiveness of the proposed feature selection method.

tion. Table 2 shows that 2-3% increase in the recognition rate was contributed by the feature selection step. Interestingly, our experiment also shows that even with Dollar's original method, by simply introducing our feature selection step, a considerable increase of performance can be achieved (see Table 2).

6. Conclusion

In this work we proposed a novel representational scheme by modelling global spatial and temporal distribution of interest points for more accurate and robust action recognition. Our method differs significantly from previous interest points based approaches in that only the global spatio-temporal distribution of the interest points are exploited. This is achieved through extracting holistic features from clouds of interest points accumulated over multiple temporal scales followed by automatic feature selection. Our approach avoids the non-trivial problem of and the current rather ad hoc approach to selecting the optimal interest point descriptor, clustering algorithm for constructing a codebook, and codebook size faced employed by existing interest points based methods. Moreover, our model is able to capture smooth motions, robust to view changes and occlusions, and with a low computation cost. Our experiments on the KTH and WEIZMANN datasets demonstrate that modelling explicitly global spatial and temporal distribution of interest points alone is highly discriminative and more robust for recognising actions under occlusion and non-rigid deformation. Ongoing work includes investigating how our global spatio-temporal distribution features can be fused with more conventional BOW features. We also aim to extend this approach to action recognition in a crowded environment.

Achnowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EP/E028594/1).

References

[1] A. Ali and J. Aggarwal. Segmentation and recognition of continuous human activity. In *IEEE Workshop on Detection and Recognition of Events in Video*, page 28, 2001.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, volume 2, pages 1395–1402, 2005.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[4] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733, 2003.

[5] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008.

[6] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *ECCV*, volume 1, pages 222–233, 2008.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 29(12):2247–2253, December 2007.

[8] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. volume 1, pages 166–173, Los Alamitos, CA, USA, 2005. IEEE Computer Society.

[9] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.

[10] J. Liu and M. Shah. Learning human actions via information maximization. In *CVPR*, 2008.

[11] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.

[12] S. Nowozin, G. H. Bakir, and K. Tsuda. Discriminative subsequence mining for action classification. In *ICCV*, pages 1–8, 2007.

[13] J. Parker. *Algorithms for Image Processing and Computer Vision*. Wiley Computer Publ., 1997.

[14] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *PAMI*, 2:1226–1238, 2005.

[15] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003.

[16] C. Rao and M. Shah. View-invariance in action recognition. In *CVPR*, volume 2, pages 316–322, 2001.

[17] S. Savarese, A. D. Pozo, J. Niebles, and L. Fei-Fei. Spatial-temporal correlations for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing*, 2008.

[18] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, volume 3, pages 32–36, 2004.

[19] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of a human action. In *ICCV*, 2005.

[20] L. Wang and D. Suter. Learning and matching of dynamic shape manifolds for human action recognition. *IEEE Transactions on Image Processing*, 16(6):1646–1661, 2007.

[21] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *CVPR*, pages I: 984–989, 2005.

[22] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia. Motion context: A new representation for human action recognition. In *ECCV*, volume 4, pages 817–829, 2008.

[23] Z. Zhao and A. Elgammal. Information theoretic key frame selection for action recognition. In *BMVC*, 2008.