Computational Visual Media
DOI 10.1007/s41095-016-0033-9 Vol. 2, No. x, month 2016, ???–???

Research Article

Saliency guided local and global descriptors for effective action recognition

Ashwan Abdulmunem[1,2](✉), **Yu-Kun Lai**[1], **and Xianfang Sun**[1]

Abstract This paper presents a novel framework for human action recognition based on salient object detection and a new combination of local and global descriptors. We first detect salient objects in video frames and only extract features for such objects. We then use a simple strategy to identify and process only those video frames that contain salient objects. Processing salient objects instead of all frames not only makes the algorithm more efficient, but more importantly also suppresses the interference of background pixels. We combine this approach with a new combination of local and global descriptors, namely 3D-SIFT and histograms of oriented optical flow (HOOF), respectively. The resulting *saliency guided 3D-SIFT–HOOF* (SGSH) feature is used along with a multi-class support vector machine (SVM) classifier for human action recognition. Experiments conducted on the standard KTH and UCF-Sports action benchmarks show that our new method outperforms the competing state-of-the-art spatiotemporal feature-based human action recognition methods.

Keywords action recognition; saliency detection; local and global descriptors; bag of visual words (BoVWs); classification

1 Introduction

Action recognition is a fundamental task and step for many problems in computer vision such as automatic visual surveillance, video retrieval, and human computer interaction. It remains a challenging research area for several reasons. Firstly, people under observation can be different in appearance, posture, and size. Secondly, moving backgrounds, occlusion, non-stationary cameras, and complex environments can impede observations. Finally, high dimensionality and low quality of video data increase the complexity and difficulty of developing efficient and robust recognition algorithms.

Therefore, extracting discriminative and informative features from video frames is challenging. Designing new methods that combine different types of features has become an important issue in action recognition. Recently, various successful approaches have been adapted from object detection and recognition in the image domain to action recognition in the video domain [1, 2]. Researchers have proposed methods based on local representations [3–5] that describe characteristics of local regions, global representations [6, 7] that describe video frame characteristics, or a combination of local and global representations [8] to improve the accuracy and benefit from both representations. Local descriptors represent a video as features extracted from a collection of patches, ideally invariant to environmental clutter, appearance change, and occlusion, and possibly to rotation and scale change as well. Global descriptors, on the other hand, treat each video frame as a whole, which is easier to implement and has lower computational costs. Combining features can take the advantages of individual features and provide a trade-off between performance and effectiveness.

To address the action recognition problem, we take a powerful, commonly used *bag of visual words* (BoVWs) pipeline and focus on the feature

1 School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 3AA, UK. E-mail: A. Abdulmunem, AbdulmunemAA@cardiff.ac.uk(✉); Y.-K. Lai, LaiY4@cardiff.ac.uk; X. Sun, SunX2@cardiff.ac.uk.

2 Department of Computer Science, School of Science, Kerbala University, Kerbala, Iraq.

Manuscript received: 2015-12-01; accepted: 2015-12-09

extraction step for performance improvement. We extract features on foreground objects identified by *saliency* and use a new combination of *local* and *global* features that provide effective complementary information (see Fig. 1). Experiments have been performed on standard datasets (KTH and UCF-Sports), which show that the proposed method outperforms use of state-of-the-art features for action recognition. The use of saliency reduces the number of feature descriptors and thus also makes the algorithm faster. More specifically, the major contributions of this paper are:

1. Each video frame contains many points of interest, making their descriptions expensive to compute. However, not all points of interest are equally important. We estimate the importance of points of interest by salient object detection, and only keep those points of interest on salient objects to perform action recognition. This helps to suppress background interference and thus makes the method more robust to background fluctuations, while at the same time reduces the running time.

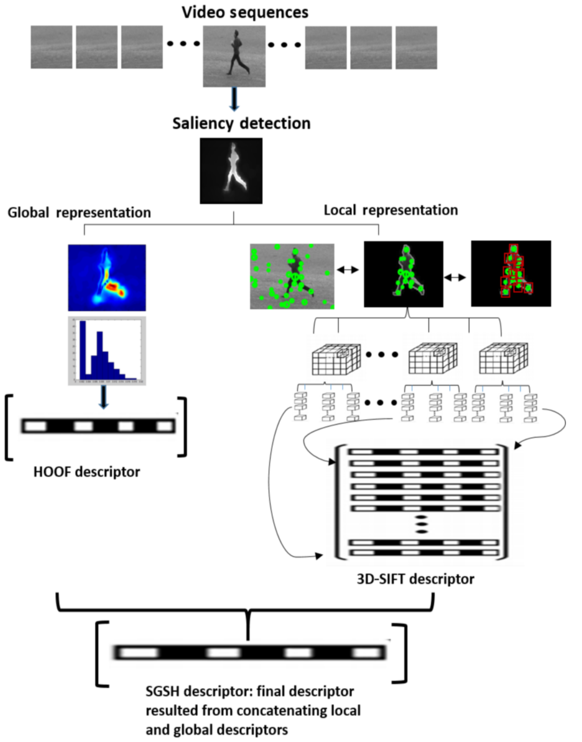


Fig. 1 Overview of our novel saliency guided feature extraction pipeline. Given a video sequence, the foreground object pixels are first identified in each frame using a saliency detection method. We then extract a new combination of local and global features guided by saliency, namely 3D-SIFT for local features and histograms of oriented optical flow (HOOF) for global features.

2. We further use a simple strategy to filter out frames that do not contain foreground objects, further improving performance and efficiency.
3. We use a novel combination of local and global descriptors, which shows good performance in action recognition.

The remaining sections of the paper are organised as follows. Section 2 discusses related work. Section 3 gives details of the proposed approach. The experimental setup and results are discussed in Section 4. Finally, conclusions are drawn in Section 5.

2 Related work

Different approaches have been introduced to address the action recognition problem. Depending on feature representations, action recognition systems can be categorised into ones based on shape and appearance-based representations [6, 9], optical-flow based representations [10, 11], and point of interest based representations [1, 2, 5]. Appearance and shape-based approaches build models to represent actions and use these models in recognition. Optical-flow approaches depend on calculating the optical flow to encode the energy of the action and represent actions as histograms of optical flow. The last representation is based on point of interest detection and local feature descriptions. As our proposed method is related to point of interest and optical-flow based approaches, we will discuss some of the most popular approaches in these two categories in the rest of this section.

Point of interest based approaches detect points of interest considered to be more informative than the others, and describe them using some feature descriptors. Many approaches have been proposed to detect points of interest. The most popular ones include a space–time points of interest detector (STIPs) [12], Harries3D (which extends the Harries detector to 3D) [13], temporal Gabor filters [14, 15], and a Hessian detector (based on the determinant of the spatiotemporal Hessian matrix).

Regarding local feature descriptors, many efforts have been made to extract and describe meaningful and robust information. Several feature descriptors have been successfully adapted from the image domain to the video domain to enhance the accuracy of human action recognition. Scovanner et al. [2] extended the SIFT descriptor [16] to the spatiotemporal domain. SIFT descriptors are invariant to changes of scale and rotation, and robust to noise. Willems et al. [3] proposed the extended SURF (ESURF) descriptor which is the generalisation of the SURF descriptor to video; it works by considering changing scales and orientations. Their evaluation however was conducted only on datasets such as KTH with a single actor and a clear recording environment.

Kläser et al. [1] represented video sequences as a 3D histogram of gradients. They extended the idea of histogram of oriented gradients (HOG) [17] from image to video to allow dense sampling of the cuboid with different scales and locations in the spatiotemporal representation. Laptev et al. [18] proposed the combined HOG/HOF descriptor which represents appearance by HOG and a local motion by histograms of flow (HOF) [11]. This descriptor is computationally expensive.

The framework of local spatiotemporal features with bag of visual words (BoVWs) has notable achievements for action recognition. Niebles and Li [9] represented video as spatiotemporal features using bag of visual words. They extracted the points of interest and clustered the features, and then modelled actions by using a probabilistic latent semantic analysis (pLSA) to localise and categorise human actions. Laptev and Lindeberg [13] recognised actions based on point of interest features. They first detected points of interest using a Hessian detector, and then described the features using scale-invariant spatiotemporal descriptors. Finally, they clustered and recognised actions based on similarity of words inside the clusters and the differences between clusters.

Moreover, BoVWs representation has become one of the most popular approaches in recent work on action recognition [5, 19–22] and shows a remarkable performance improvement on some benchmark datasets. Wang et al. [21] recognised actions using a BoVW framework with an SVM classifier. They represented the video by a combination of several descriptors: histograms of oriented gradient to describe the appearance, histograms of optical flow (motion) and trajectories to describe the shape. Moreover they introduced a descriptor based on motion boundary histograms (MBH) which rely on differential optical flow.

Recently, Zhang et al. [5] introduced a 3D feature descriptor called simplex-based orientation decomposition (SOD) with BoVW to recognise actions. The SOD descriptor is based on decomposing the visual cue orientations into three angles and transforming the decomposed angles into the simplex space. The histograms of 3D visual cues are then obtained in the simplex space which are then used to form the final feature vectors for classification.

Oikonomopouls et al. [23] adapted the idea of saliency region selection in spatial images to spatiotemporal video space. Salient points are detected by measuring changes in the information content of the set of pixels in cylindrical spatiotemporal neighbourhoods at different scales. They used sparse representation of a human action as a set of spatiotemporal salient points that correspond to activity-variation peaks to recognise the action. Their method directly uses saliency information as features for action recognition, whereas we use saliency information to guide more general feature descriptors.

3 Proposed approach

In this section we describe our proposed approach for action recognition. The pipeline is illustrated in Fig. 2, which contains the following four main steps. The first step is saliency guided feature extraction, where the salient objects are detected and only points of interest on the objects are used. With saliency as guidance, local and global features are then extracted to encode video information. In the training step, features extracted from the training set are clustered to generate visual words. Histograms based on occurrences of visual words in the training set are used as features to train classifiers. Finally, a multi-class SVM classifier is used to achieve action recognition. The following subsections explain each step in detail.

3.1 Saliency guided feature extraction

The first step of our pipeline is to perform saliency guided feature extraction (SGFE) in video frames. This provides a fast solution that addresses several key aspects related to action recognition. Firstly, it detects the region of interest (ROI) and attention-grabbing objects in a scene. Secondly, it selects the informative and robust keypoints in the frames.

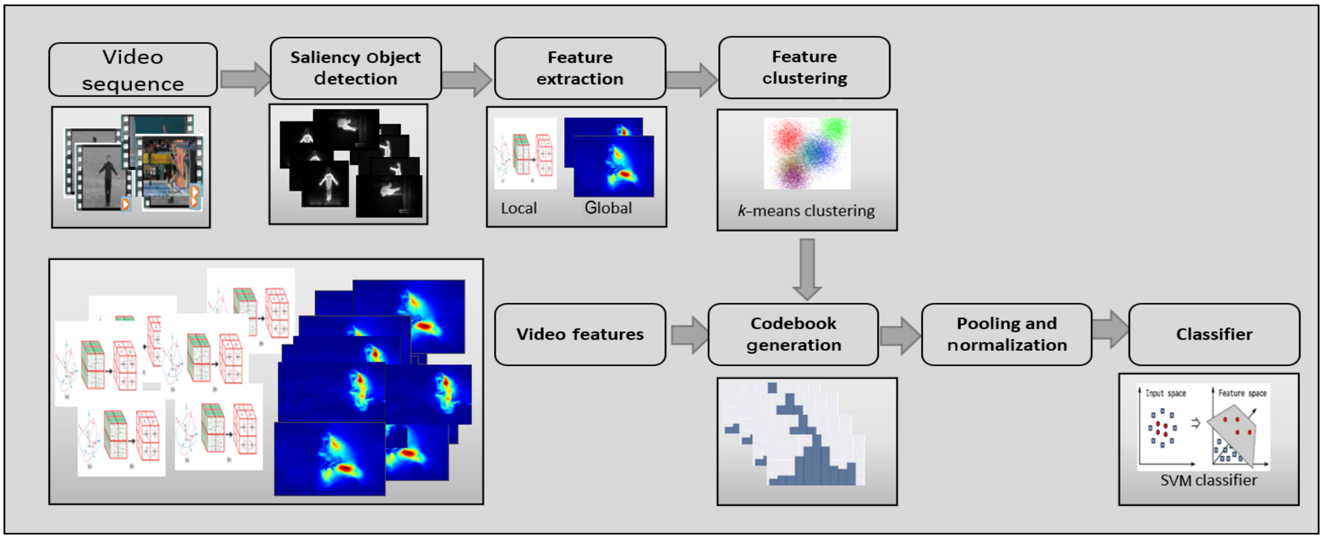


Fig. 2 The pipeline for obtaining a bag of visual words (BoVWs) representation for action recognition. It contains five main steps: (i) saliency guided feature extraction, (ii) feature clustering, (iii) codebook dictionary generation, (iv) pooling and normalisation, and (v) classification.

Finally, it reduces the time required to encode each video frame. Saliency detection research has largely been based on images. We use a state-of-the-art image-based algorithm [24] and apply this to each video frame. The main idea of this algorithm is to combine colour and pattern distinctness. It is inspired by the fact that the neighbouring pixels of each salient object are distinct in both colour and pattern. Colour detection is performed by segmenting a video frame into regions and then determining which region is distinct in colour. The colour distinctness of a region is defined as the sum of L_2 distances from all other regions in the colour space.

Pattern distinctness is determined by firstly extracting all 9×9 patches and computing the average patch. Principal component analysis (PCA) is then applied to the collection of patches. After that the pattern distance of a patch is defined as the L_1 distance between the patch and the average patch, calculated in PCA coordinates. Doing so takes into account not only the difference between a patch and the average patch, but also the distribution of patches. Unusual patches based on the distribution have a high pattern distinctness. Because objects are more likely to be in the centre of the frame, a Gaussian map surrounding the centre of the frame is also generated. The final saliency space map $S(p_x)$ is the product of the colour distinctness map, patch distinctness map, and the Gaussian map.

After salient object detection, a binary image is generated by thresholding (threshold 0.2 is used in our experiments) and used as a mask to extract foreground object from the background. Figure 3 presents examples of salient object detection for some actions in both the *KTH* and *UCF-Sports* datasets. Saliency detection works well for both datasets. Note that we have found applying the image based saliency detection technique to individual frames works very well for these datasets, including the UCF-Sports dataset with complex background. As will be explained later in the paper, histogram-based features are used for classification, which makes the system more robust to inaccuracies of saliency detection in individual frames.

As we will show later, this step improves the performance substantially by selecting only the points of interest which are detected on objects and discarding others in the background. Furthermore, we use video frame selection, keeping only those video frames containing foreground subjects for further processing. For frames without foreground subjects, the saliency detector tends to classify background areas as salient regions. An example illustrating this is shown in Fig. 4. Since the background usually covers more pixels than the foreground, we retain for further processing frames with fewer than half of the pixels being classified as salient, and discard the remaining frames. This simple heuristic works well for all the datasets tested

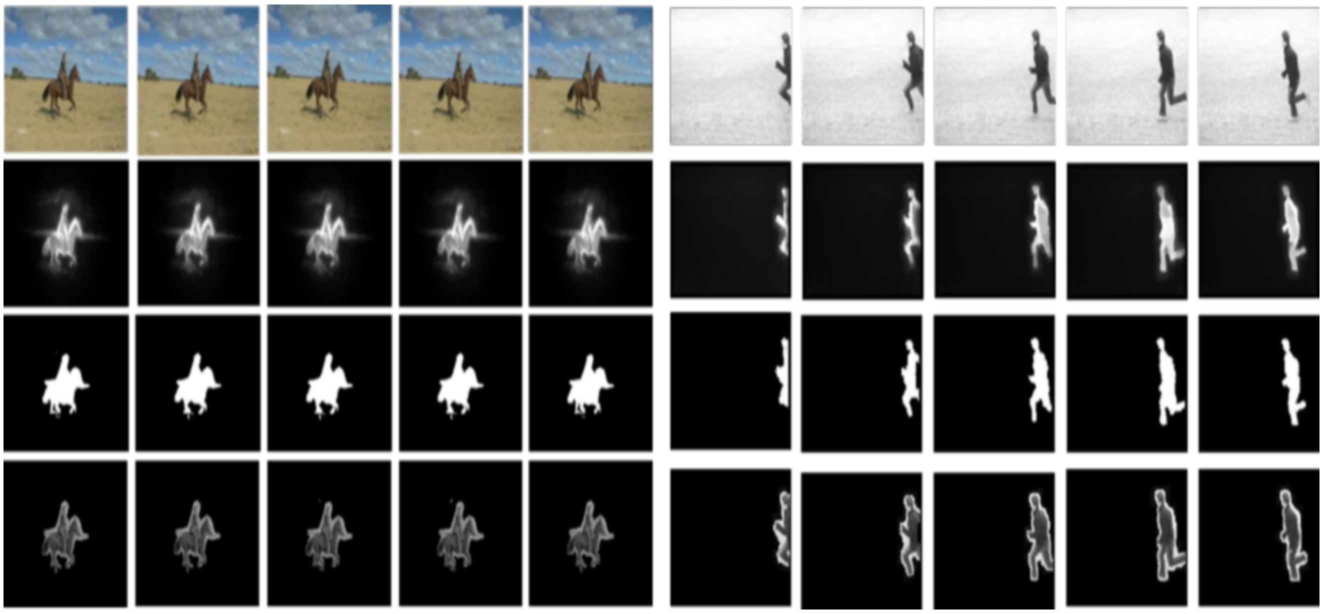


Fig. 3 Salient object detection. First row: original video frames. Second row: results of saliency detection. Third row: binary images for the processed frames. Fourth row: foreground objects in video frames. The left five columns give an example from UCF-Sports (horse-riding) and the right five columns give an example from the KTH dataset (running).

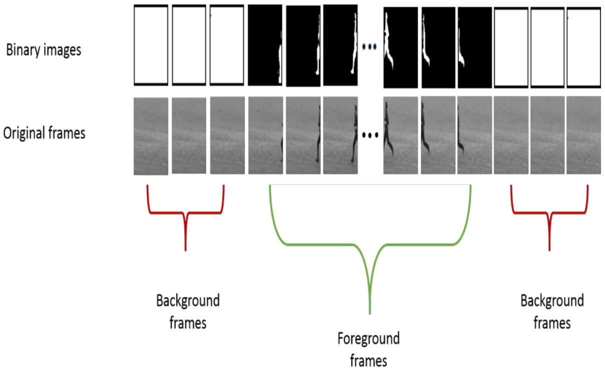


Fig. 4 Proposed video frame selection.

in this work.

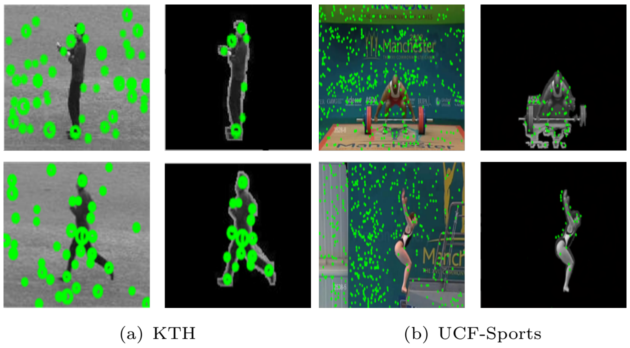
For local representation, we need to first detect points of interest in video frames. In this step, a common approach is to use the Laplacian of Gaussian (LoG) as the response function. We use Lowe's approach to extract the points of interest [16]. An approximation of the LoG is used based on the difference of the image smoothed at different scales. The response function is

$$D = (g(.; k\sigma) - g(.; \sigma)) * I = L(.; k\sigma) - L(.; \sigma) \quad (1)$$

where k is a parameter which controls the accuracy of the approximation, g is a 2D Gaussian kernel with a given standard deviation and $L(.; \sigma) = g(.; \sigma) * I$. We select only the points of interest detected on salient objects. Consequently, we process the most important points in the video frames, which carry robust information of an action. All points detected on the background are discarded. The motivation for this is that the salient points of interest are precisely those that maximise the discriminability of actions. Figure 5 shows the difference between the points of interest detected before and after applying salient object detection for examples from both datasets *KTH* (*boxing*, *running*) and *UCF-Sports* (*lifting*, *diving*).

3.2 Feature description

We extract two types of descriptors: local and global descriptors represented by *3D-scale invariant feature transform* (3D-SIFT) and *histogram of oriented optical flow* (HOOF), respectively (see



(a) KTH (b) UCF-Sports

Fig. 5 Point of interest detection on KTH and UCF-Sports datasets: the first and third columns are the original frames, and the second and fourth columns are frames after salient object detection.

Fig. 1). Local representation provides detailed information insensitive to global transformations. The scale invariant feature transform (SIFT) descriptor is one of the most popular local representations due to its invariance to camera movement, and is robust to noise and scaling. We detect points of interest using image-based SIFT for each video frame, and use 3D-SIFT descriptors [2, 8] to represent local features of points of interest, owing to the fact that video frames have a spatiotemporal domain.

Motion representation as a global descriptor is particularly useful in action representation due to its low computational cost and capability of capturing global motions. In our approach, we describe the motion using the HOOF descriptor [10] for each video frame. For optical-flow calculation, we use Brox's method [25] as shown in Fig. 6.

3.3 Classification

The actions are represented by a bag of visual words model with concatenation of local spatiotemporal and global features. For normalization, the sum of all features for each video is set to 1. For classification, we use a multi-class support vector machine (SVM) with a radial basis function (RBF) kernel [26]. A bag of visual words approach is used to encode the videos. In clustering we use k-means algorithm to generate the vocabulary of visual words. Each feature vector is mapped to the closest visual word and a video is then represented as a frequency histogram of the visual words.

4 Experimental results

In this section, we describe the datasets used and the experimental results using our approach.

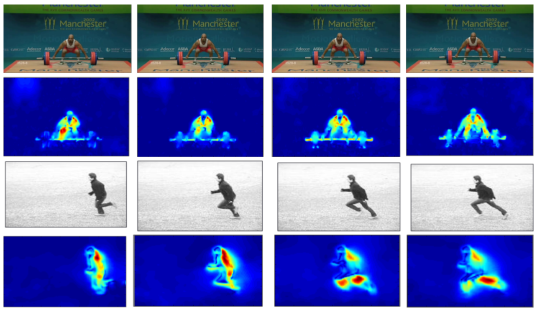


Fig. 6 Optical-flow calculation using Brox's method: the first two rows show an example from the UCF-Sports dataset (lifting) and the last two rows for the KTH dataset (running).

4.1 Datasets

To evaluate the performance of our approach, we conducted experiments on the KTH and UCF-Sports datasets. The KTH dataset consists of 6 actions (*boxing*, *clapping*, *waving*, *jogging*, *walking*, and *running*) which were recorded under controlled settings with approximately static cameras, a clear environment, and in different scenarios (outdoors, outdoors at a different scale, outdoors with different clothes, and indoors).

The UCF-Sports dataset contains 10 sport actions *diving*, *golf swinging*, *kicking*, *lifting*, *horseback riding*, *running*, *skating*, *swinging*, *walking*. The UCF-Sports dataset has large intra-class variation with real world recording environment settings.

A standard test setup was used (training/test separation for KTH and leave-one-out testing for UCF-Sports) to allow fair comparison with prior work.

4.2 Parameters

The patch size for the SIFT descriptor is a cube of size 8 × 8 × 8; each cube is divided into sub-cubes of size 4 × 4 × 4. For each sub-cube an orientation histogram with 8 bins is produced, so we have 24 bins for each sub-cube, and for the whole cube all these sub-cube histograms are combined to form a 192 (= 24 × 8) dimensional feature vector, which is the 3D-SIFT feature descriptor. For the HOOF descriptor, each video frame is represented by a feature vector with 150 bins. In our experiments, vocabularies are constructed with k-means clustering with 1000 visual words for 3D-SIFT and 2000 for HOOF. Grid search with 5-fold cross validation was used to optimise SVM kernel parameters.

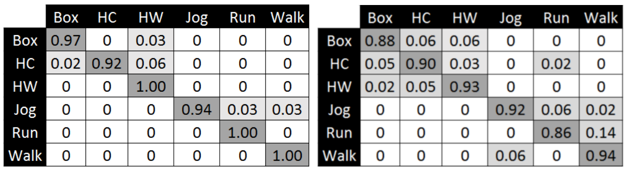
4.3 Results

Table 1 shows our experimental results on the KTH and UCF-Sports datasets for cases with and without saliency guidance (i.e., *SGSH* and *SH* descriptors) respectively.

From the table, we can see that the SGSH descriptor increases the accuracy by 6% for the KTH dataset and 5.6% for the UCF-Sprots dataset. Figures 7 and 8 show the confusion matrices for the KTH and UCF-Sports datasets, respectively. It can be clearly seen that, using the SGSH descriptor, the actions *hand waving*, *running*, and *walking* in the

Table 1 Action recognition with and without saliency guidance for the combined 3D-SIFT and HOOF descriptors (SGSH and SH, respectively)

Dataset	Descriptor	Accuracy (%)
KTH	SGSH	**97.2**
	SH	91.2
UCF-Sports	SGSH	**90.9**
	SH	85.3



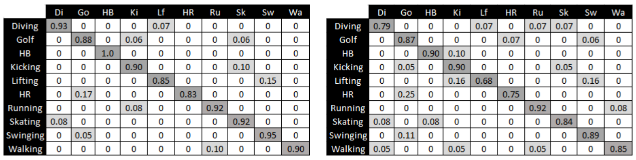
	Box	HC	HW	Jog	Run	Walk
Box	0.97	0	0.03	0	0	0
HC	0.02	0.92	0.06	0	0	0
HW	0	0	1.00	0	0	0
Jog	0	0	0	0.94	0.03	0.03
Run	0	0	0	0	1.00	0
Walk	0	0	0	0	0	1.00

(a)

	Box	HC	HW	Jog	Run	Walk
Box	0.88	0.06	0.06	0	0	0
HC	0.05	0.90	0.03	0	0.02	0
HW	0.02	0.05	0.93	0	0	0
Jog	0	0	0	0.92	0.06	0.02
Run	0	0	0	0	0.86	0.14
Walk	0	0	0	0.06	0	0.94

(b)

Fig. 7 Confusion matrices for the KTH dataset (HC: hand clapping, HW: hand waving): (a) SGSH and (b) SH.



	Di	Go	HB	Ki	Lf	HR	Ru	Sk	Sw	Wa
Diving	0.93	0	0	0	0.07	0	0	0	0	0
Golf	0	0.88	0	0.06	0	0	0	0.06	0	0
HB	0	0	1.0	0	0	0	0	0	0	0
Kicking	0	0	0	0.90	0	0	0	0.10	0	0
Lifting	0	0	0	0	0.85	0	0	0	0.15	0
HR	0	0.17	0	0	0	0.83	0	0	0	0
Running	0	0	0	0.08	0	0	0.92	0	0	0
Skating	0.08	0	0	0	0	0	0	0.92	0	0
Swinging	0	0.05	0	0	0	0	0	0	0.95	0
Walking	0	0	0	0	0	0	0.10	0	0	0.90

(a)

	Di	Go	HB	Ki	Lf	HR	Ru	Sk	Sw	Wa
Diving	0.79	0	0	0	0.07	0	0.07	0.07	0	0
Golf	0	0.87	0	0	0	0.07	0	0	0.06	0
HB	0	0	0.90	0.10	0	0	0	0	0	0
Kicking	0	0.05	0	0.90	0	0	0	0.05	0	0
Lifting	0	0	0	0.16	0.68	0	0	0	0.16	0
HR	0	0.25	0	0	0	0.75	0	0	0	0
Running	0	0	0	0	0	0	0.92	0	0	0.08
Skating	0.08	0	0.08	0	0	0	0	0.84	0	0
Swinging	0	0.11	0	0	0	0	0	0	0.89	0
Walking	0.05	0	0	0.05	0	0	0.05	0	0	0.85

(b)

Fig. 8 Confusion matrices for the UCF-Sports dataset (HB: high bar swinging, HR: horse riding): (a) SGSH and (b) SH.

KTH dataset are recognised fully correctly, while using the SH descriptor, confusion exists between these actions and others. For example 6% of *walking* actions were recognised as *jogging*. Compared to the existing methods, our approach shows an effective increase in performance (see Tables 2 and 3).

4.4 Running time

From a computational cost point of view, SGSH reduces the time required to process the points of interest by reducing the number of points of interest detected in each video frame and selecting only the informative frames, as shown in Table 4 for the *boxing* action as an example. The number of points of interest reduces substantially with saliency guidance. Moreover, after video frame selection, the number of processed frames is also reduced significantly, as shown in Table 5 for the *running* action as an example. In general, our current unoptimised implementation on a 2.5 GHz Windows 8 workstation takes a mean CPU-time of 0.09 seconds to process a point of interest. A whole frame takes 1.9 seconds for the 3D-SIFT descriptor and 0.4 seconds for the HOOF descriptor. On average, a 2–5 times speedup is provided by saliency guidance due to the reduced number of feature points in each frame and the reduced number of frames to be detected.

Table 2 Recognition accuracy comparisons on the KTH dataset

Method on KTH	Accuracy (%)
Al Ghamdi et al. [27]	90.7
Liu et al. [28]	91.3
Iosifidis et al. [29]	92.1
Baumann et al. [30]	92.1
Kläser [31]	92.6
Ji et al. [32]	93.1
Wang et al. [33]	94.2
Wu et al. [34]	94.5
Raptis and Soatto [35]	94.8
Zhang et al. [5]	94.8
Wang et al. [21]	95.0
Yuan et al. [4]	95.4
SGSH	**97.2**

Table 3 Recognition accuracy comparisons on the UCF-Sports dataset

Method on UCF-Sports	Accuracy (%)
Raptis et al. [36]	79.4
Ma et al. [37]	81.7
Kläser [31]	85.0
Everts et al. [38]	85.6
Le et al. [39]	86.5
Yuan et al. [4]	87.3
Zhang et al. [5]	87.5
Wang et al. [21]	88.0
Wang et al. [33]	88.2
Ma et al. [19]	89.4
SGSH	**90.9**

5 Conclusions

In this paper, we have proposed a novel video feature extraction method based on saliency detection with a new combination of local and global descriptors. We detect salient objects in each video frame and process only the points of interest in these objects. We also use video frame selection to discard all frames without salient objects. Experiments show that the proposed method gives a significant improvement for the action recognition for both datasets (see Tables 2 and 3). Our method outperforms state-of-the-art features using BoVW based classification methods. The idea of using saliency guidance to improve action recognition is

Table 4 Average number of points of interest with and without saliency guidance (using the *boxing* video as an example). The first column is the average number of keypoints detected on the video frames. The second is the average number of keypoints detected on the object

Points of interest/frame	After SGFE
41	24
47	23
39	19
43	26
52	31

Table 5 Results of the proposed video frame selection approach (using *running* as an example). First column: duration of the video. Second column: number of frames in the video. Third column: number of the frames which contain the foreground object (object on-screen)

Duration (s)	Number of frames	Object on-screen
00:00:20	500	165
00:00:13	345	122
00:00:26	666	336
00:00:22	570	181
00:00:14	248	133

general and in the future we hope to investigate combining it with alternative features as well as to consider its use in other recognition applications.

Acknowledgements

This research is funded by Iraqi Ministry of Higher Education and Scientific Research (MHESR).

References

[1] Kläser, A.; Marszalek, M.; Schmid, C. A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of British Machine Vision Conference, 995–1004, 2008.

[2] Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia, 357–360, 2007.

[3] Willems, G.; Tuytelaars, T.; Van Gool, L. An efficient dense and scale-invariant spatiotemporal interest point detector. In: *Lecture Notes in Computer Science, Vol. 5303*. Forsyth, D.; Torr, P.; Zisserman, A. Eds. Springer Berlin Heidelberg, 650–663, 2008.

[4] Yuan, C.; Li, X.; Hu, W.; Ling, H.; Maybank, S. 3D R transform on spatiotemporal interest points for action recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 724–730, 2013.

[5] Zhang, H.; Zhou, W.; Reardon, C.; Parker, L. Simplex-based 3D spatio-temporal feature description for action recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2067–2074, 2014.

[6] Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 35, No. 1, 221–231, 2013.

[7] Taylor, G. W.; Fergus, R.; LeCun, Y.; Bregler, C. Convolutional learning of spatiotemporal features. In: Proceedings of the 11th European Conference on Computer Vision: Part VI, 140–153, 2010.

[8] Sun, X.; Chen, M.; Hauptmann, A. Action recognition via local descriptors and holistic features. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 58–65, 2009.

[9] Niebles, J. C.; Li, F.-F. A hierarchical model of shape and appearance for human action classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2007.

[10] Chaudhry, R.; Ravichandran, A.; Hager, G.; Vidal, R. Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1932–1939, 2009.

[11] Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. In: Proceedings of the 9th European Conference on Computer Vision, Vol. 2, 428–441, 2006.

[12] Laptev, I. On space–time interest points. *International Journal of Computer Vision* Vol. 64, Nos. 2–3, 107–123, 2005.

[13] Laptev, I.; Lindeberg, T. Space–time interest points. In: Proceedings of the 9th IEEE International Conference on Computer Vision, 432–439, 2003.

[14] Bregonzio, M.; Gong, S.; Xiang, T. Recognising action as clouds of space–time interest points. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1948–1955, 2009.

[15] Dollar, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior recognition via sparse spatio-temporal features. In: Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 65–72, 2005.

[16] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* Vol. 60, No. 2, 91–110, 2004.

[17] Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 886–893, 2005.

[18] Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2008.

[19] Ma, S.; Sigal, L.; Sclaroff, S. Space–time tree ensemble for action recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 5024–5032, 2015.

[20] Qu, T.; Liu, Y.; Li, J.; Wu, M. Action recognition using multi-layer topographic independent component analysis. *Journal of Information & Computational Science* Vol. 12, No. 9, 3537–3546, 2015.

[21] Wang, H.; Klaser, A.; Schmid, C.; Liu, C.-L. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision* Vol. 103, No. 1, 60–79, 2013.

[22] Wu, J.; Zhang, Y.; Lin, W. Towards good practices for action video encoding. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2577–2584, 2014.

[23] Oikonomopoulos, A.; Patras, I.; Pantic, M. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* Vol. 36, No. 3, 710–719, 2005.

[24] Margolin, R.; Tal, A.; Zelnik-Manor, L. What makes a patch distinct? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1139–1146, 2013.

[25] Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In: Proceedings of the 8th European Conference on Computer Vision, Springer LNCS 3024. Pajdla, T.; Matas, J. Eds. Springer-Verlag Berlin Heidelberg, Vol. 4, 25–36, 2004.

[26] Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* Vol. 2, No. 3, Article No. 27, 2011.

[27] Al Ghamdi, M.; Zhang, L.; Gotoh, Y. Spatiotemporal SIFT and its application to human action classification. In: *Lecture Notes in Computer Science, Vol. 7583.* Fusiello, A.; Murino, V.; Cucchiara, R. Eds. Springer Berlin Heidelberg, 301–310, 2012.

[28] Liu, J.; Kuipers, B.; Savarese, S. Recognizing human actions by attributes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 3337–3344, 2011.

[29] Iosifidis, A.; Tefas, A.; Pitas, I. Discriminant bag of words based representation for human action recognition. *Pattern Recognition Letters* Vol. 49, 185–192, 2014.

[30] Baumann, F.; Ehlers, A.; Rosenhahn, B.; Liao, J. Recognizing human actions using novel space–time volume binary patterns. *Neurocomputing* Vol. 173, No. P1, 54–63, 2016.

[31] Kläser, A. Learning human actions in video. Ph.D. Thesis. Université de Grenoble, 2010.

[32] Ji, Y.; Shimada, A.; Nagahara, H.; Taniguchi, R.-i. A compact descriptor CHOG3D and its application in human action recognition. *IEEJ Transactions on Electrical and Electronic Engineering* Vol. 8, No. 1, 69–77, 2013.

[33] Wang, H.; Klaser, A.; Schmid, C.; Liu, C.-L. Action recognition by dense trajectories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 3169–3176, 2011.

[34] Wu, X.; Xu, D.; Duan, L.; Luo, J. Action recognition using context and appearance distribution features. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 489–496, 2011.

[35] Raptis, M.; Soatto, S. Tracklet descriptors for action modeling and video analysis. In: Proceedings of the 11th European Conference on Computer vision: Part I, 577–590, 2010.

[36] Raptis, M.; Kokkinos, I.; Soatto, S. Discovering discriminative action parts from mid-level video representations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1242–1249, 2012.

[37] Ma, S.; Zhang, J.; Ikizler-Cinbis, N.; Sclaroff, S. Action recognition and localization by hierarchical space–time segments. In: Proceedings of IEEE International Conference on Computer Vision, 2744–2751, 2013.

[38] Everts, I.; van Gemert, J. C.; Gevers, T. Evaluation of color STIPs for human action recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2850–2857, 2013.

[39] Le, Q. V.; Zou, W. Y.; Yeung, S. Y.; Ng, A. Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 3361–3368, 2011.

Ashwan Abdulmunem received her B.S. and M.S. degrees in computer science from Babylon University, Babylon, Iraq. She is a lecturer at Kerbala University, and currently, she is a Ph.D. student at the School of Computer Science and Informatics, Cardiff University. Her research interests include computer vision, video processing, pattern recognition, and artificial intelligence.

Yu-Kun Lai received his bachelor and Ph.D. degrees in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a lecturer of visual computing at the School of Computer Science and Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing, and computer vision. He is on the Editorial Board of *The Visual Computer*.

Xianfang Sun received his Ph.D. degree in control theory and its applications from the Institute of Automation, Chinese Academy of Sciences. He is a senior lecturer at Cardiff University. His research interests include computer vision and graphics, pattern recognition and artificial intelligence, and system identification and control.

Other papers from this open access journal are available free of charge from http://www.springer.com/journal/41095. To submit a manuscript, please go to https://www.editorialmanager.com/cvmj.