Chapter 58
Human Action Recognition using Salient Region Detection in Complex Scenes

Zhong Zhang, Shuang Liu, Shuaiqi Liu, Liang Han, Yunxue Shao, and Wen Zhou

Abstract Although the methods based on spatio-temporal interest points have shown promising results for human action recognition, they are not robust in complex scenes especially background clutter, camera motion, occlusions and illumination variations. In this paper, we propose a novel method to classify human actions in complex scenes. We suppress the false detection interest points by detecting salient regions. Furthermore, we encode the features according to their spatio-temporal relationship. Our method is verified on two challenging databases (UCF sports and YouTube), and the experimental results demonstrate that our method achieves better results than previous methods in human action recognition.

Keywords Human action recognition • Salient region detection • Complex scenes

58.1 Introduction

Automatically recognizing human actions is receiving increasing attention due to its wide range of applications such as video retrieval, human-computer interaction and activity monitoring. A large number of methods [1, 2] for humane action recognition have been proposed, ranging from trajectory-based methods [3] and local descriptor-based methods [4] to attribute-based method [5, 6].

Z. Zhang • S. Liu (✉) • S. Liu • L. Han
College of Electron and Communication Engineering, Tianjin Normal University, Tianjin, China
e-mail: zhong.zhang8848@gmail.com; shuangliu.tjnu@gmail.com; shdkj-1918@163.com; hanliang@mail.tjnu.edu.cn

Y. Shao
College of Computer Science, Inner Mongolia University, Inner Mongolia, China
e-mail: csshyx@imu.edu.cn

W. Zhou
The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: wen.zhou@ia.ac.cn

J. Mu et al. (eds.), *The Proceedings of the Third International Conference on Communications, Signal Processing, and Systems*, Lecture Notes in Electrical Engineering 322, DOI 10.1007/978-3-319-08991-1_58

However, most of these previous approaches for human action recognition are constrained to well-controlled environments and fail to achieve desired results in complex scenes. Human action recognition in complex scenes is an extremely difficult task, due to several challenges, like background clutter, camera motion, occlusions and illumination variations. To address these challenges, several methods, such as tree-based template matching [7], tensor canonical correlation [8] and prototype based action matching [9] are proposed. Most of these methods are complex, time consuming and preprocessing requirement, such as segmentation, tree data structure building, target tracking or background subtraction. Other methods [10–12] for human action recognition in complex scenes apply spatio-temporal interest point detectors and local descriptors to characterize and encode the action video, which demand less or no preprocessing. Thus, this kind of methods achieve promising recognition accuracy. However, interest points are usually false detection in uncontrolled environments and Fig. 58.1 shows the result of interest point detection in complex scenes. We can see that the interest points outside the green rectangle are invalid owing to the actor inside the green rectangle.

In this paper, we propose a novel method to classify human actions in complex scenes. As is well-known, interest points inside or around the actor are beneficial to classification. Therefore, we utilize salient regions to select the interest points. Concretely, we reserve the interest points with high salient values, while we discard the interest points with low salient values. After selecting interest points, we apply CLC coding strategy [13] to consider the spatial and temporal relationship among interest points. Finally, we train the classification model using SVM.

The rest of this paper is organized as follows. Section 58.2 introduces the proposed method in detail. Section 58.3 demonstrates that our experimental results are more accurate than the state-of-the-art methods on UCF sports dataset and YouTube dataset. Finally, in section "Conclusion" we conclude this paper.

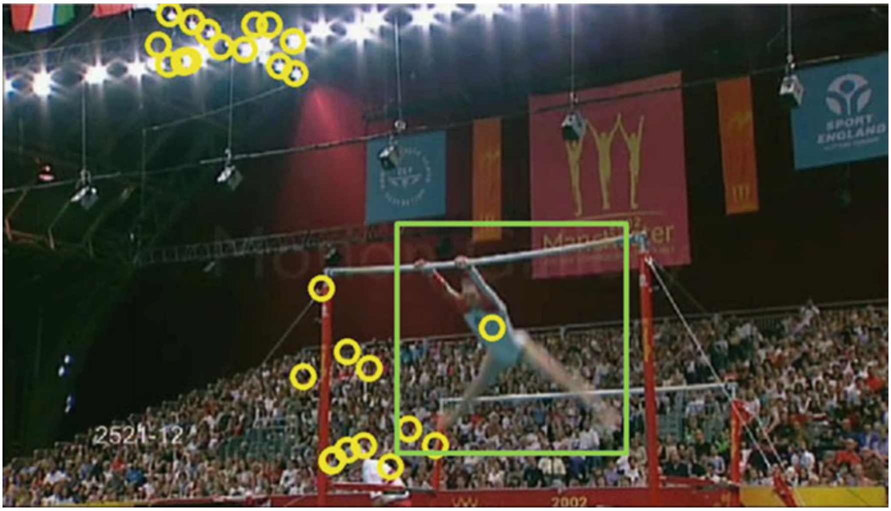


Fig. 58.1 Detecting interest points in complex scenes

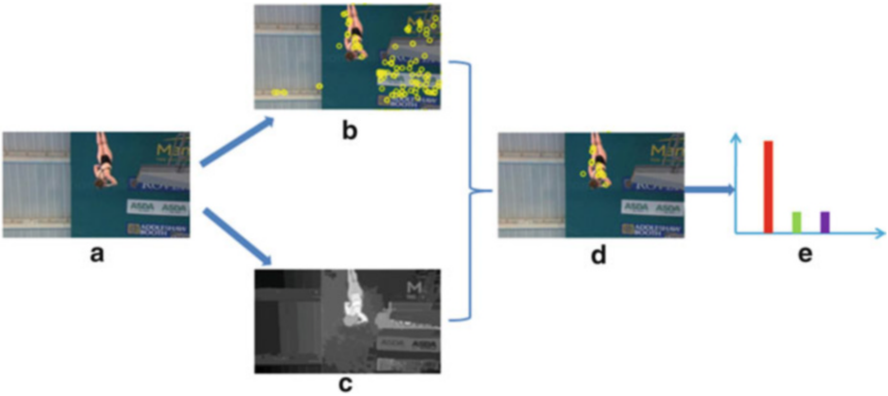


Fig. 58.2 Flowchart of the proposed method. (**a**) Original action video; (**b**) detecting spatio-temporal interest points; (**c**) detecting salient regions; (**d**) selecting significative spatio-temporal interest points; (**e**) generating feature histograms

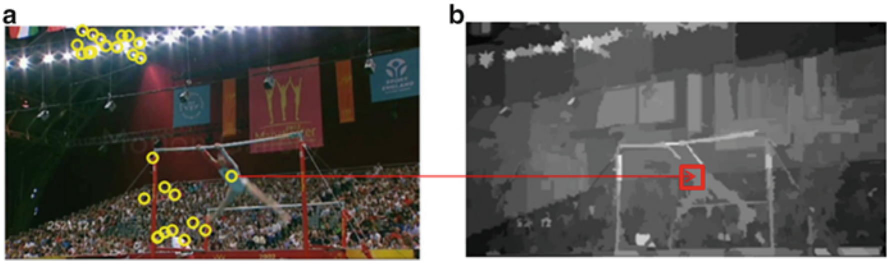


Fig. 58.3 Selective interest points

58.2 Approach

58.2.1 Method Overview

The proposed method consists of four stages: (a) detecting spatio-temporal interest points for each action video; (b) detecting salient regions for each frame of an action video; (c) selecting significative spatio-temporal interest points according salient regions; (d) generating feature histograms and training classifier. The flowchart of the proposed method is shown in Fig. 58.2 and the corresponding detailed description will be presented in the following sections.

58.2.2 Detection of Spatio-Temporal Interest Points

To detect interest points, we first use Harris3D corner detector [14] for each action video as shown in Fig. 58.2b. For each interest point, we characterize the local appearance using histogram-of-gradients (HOG) and histogram-of-optical-flow (HOF) [15]. As a result, we obtain a set of interest points for an action video, $v = (\mathbf{x}_i, \mathbf{s}_i)_{i=1,\ldots,N}$, where N is the number of interest points for the action video v, $\mathbf{x}_i$ indicates the feature vector of the i-th interest point and $\mathbf{s}_i$ is the location of the i-th interest point. Here $\mathbf{s}_i = (x, y, t)$ where x, y and t are horizontal, vertical, and temporal coordinates respectively.

58.2.3 Salient Region Detection

We detect the salient regions for each frame by using region contrast (RC) [16]. First, we segment the frame into regions using a segmentation method based on graph cut [17]. Then, the color histogram is builded for each region. For each region a_k, we calculate the salient value by comparing its color contract with all other regions:

$$S(a_k) = \sum_{a_k \neq a_i} exp(D_s(a_k, a_i)/\sigma^2) w(a_i) D_r(a_k, a_i) \tag{58.1}$$

where $w(a_i)$ is the weight of region a_i, D_r represent the color distance between the two regions, D_s is the spatial distance between two regions, and σ controls the strength of spatial weight. From the above equation, we can see that all the pixels in one region share the same salient value. The number of pixels in a_i indicates the weight w_i. The spatial distance is defined as the Euclidean distance between the centroids of regions. The color distance between two regions a_1 and a_2 is defined as:

$$D_r(a_1, a_2) = \Sigma_{i=1}^{n_1} \Sigma_{j=1}^{n_2} f(c_{1,i}) f(c_{2,j}) D(c_{1,i}, c_{2,j}) \tag{58.2}$$

where $f(c_{k,i})$ is the probability of the i-th color $c_{k,i}$ among all n_k colors in the k-th region a_k, $k = 1, 2$. The result of salient region detection is shown in Fig. 58.2c.

58.2.4 Selective Interest Points

Since human actions are usually recorded in complex scenes, for example cluttered background, illumination variations, camera motion and occluded bodies, there are a lot of noise interest points. The location of these noise interest points are usually in the background, and therefore they are injurious to classification. To address this problem, we apply the salient regions to suppress these noise interest points. For each interest point $(\mathbf{x}, \mathbf{s})$, we compute the maximum salient value around its space location:

$$S_m = \max_{(x,y)\in R_l} S((x,y)) \tag{58.3}$$

where R_l is the local region around the interest point (see the red rectangle in Fig. 58.3b) and $S((x,y)$ is the salient value at (x,y). If $S_m > T_s$, we reserve this interest point due to its location in salient region; otherwise we discard this interest point. Here T_s is the salient threshold value. The result of selective interest points is illustrated in Fig. 58.2d.

58.2.5 Feature Coding

After selecting the interest points, we cluster all the interest points from all the action videos by using k-means clustering algorithm and generate the dictionary. Then, we utilize CLC [13] strategy to code the features, which not only considers the spatio-temporal relationships among interest points, but also alleviates the quantization error by using linear coding. Afterwards, each action video is represented by a histogram (see Fig. 58.2e). Finally, we use these feature histograms to train a multi-class SVM.

58.3 Experimental Results

To evaluate our proposed method for human action recognition, we conduct a series of experiments on two publicly available human action datasets: UCF Sports dataset [18] and YouTube action dataset [19]. These two datasets are challenging because these action videos are recorded in realistic scenes and suffer from cluttered background, illumination variations, camera motion and so on. The codebook is constructed by k-means algorithm and the number of codebook is empirically set to 4,000 [20]. The salient threshold value T_s is set to 180 and R_l is set to 15.

We present the results of selective interest points in Fig. 58.4. The first row is the original detected interest points, the second row is the salient regions, and the last row is the results of suppressing the noise interest points. From Fig. 58.3, we can see that most of interest points on background are discarded and the interest points on human are reserved which is beneficial to the subsequent classification. Next, we will objectively evaluate our proposed method on UCF Sports dataset and YouTube action dataset.

The UCF sports dataset [18] contains ten different types of sports action: swinging (on the pommel horse and on the floor), diving, kicking (a ball), weight-lifting, horse-riding, running, skateboarding, swinging (at the high bar), golf swinging and walking. The dataset consists of 150 real videos with a wide range of viewpoints and scene backgrounds. In order to increase the amount of

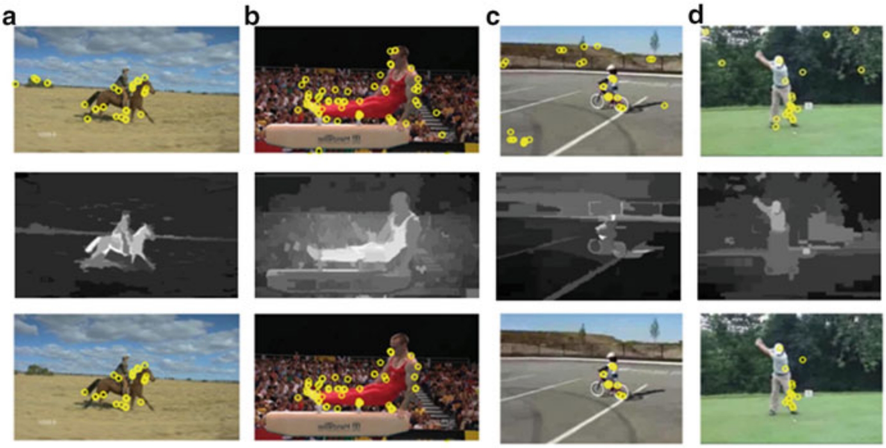


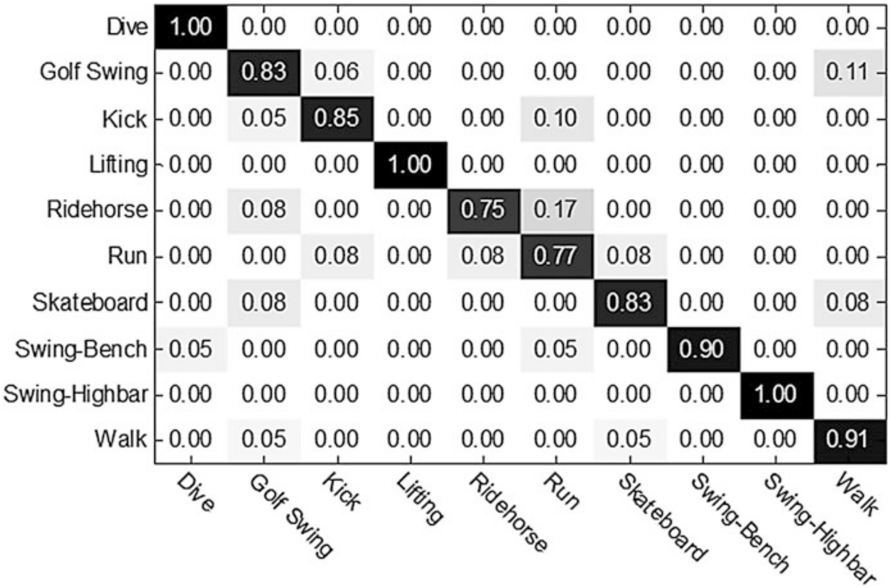
Fig. 58.4 Performance of the selective interest points in complex scenes. The action videos (**a**, **b**) are from UCF Sports dataset and the action videos (**c**, **d**) are from YouTube action dataset

Table 58.1 Recognition results of different methods on the UCF Sports dataset

Method	Accuracy (%)
Sullivan et al. [18]	69.2
Wang et al. [20]	85.6
Kovashka et al. [21]	87.27
Le et al. [22]	86.5
Zhang et al. [13]	87.33
Ours	**88.0**

training samples, we extend the dataset by adding a horizontally flipped version of each video sequence to the dataset as suggested in [20]. Table 58.1 compares our method with the other excellent methods, in which we can see that our method achieves the highest recognition accuracy of 88 %. Figure 58.5 shows the confusion table of recognition results on UCF sports dataset. From this figure, "horse-riding" are prone to be misclassified into "running" due to their similar appearance.

The YouTube dataset [19] is a collection of 1,168 complex and challenging YouTube videos of 11 human actions categories: basketball shooting, volleyball spiking, trampoline jumping, soccer juggling, horseback riding, cycling, diving, swinging, golf swinging, tennis swinging and walking (with a dog). The dataset has the following properties: a mix of steady cameras and shaky cameras, cluttered background, low resolution, and variation in object scale, viewpoint and illumination. Our method achieves 88.65 % recognition accuracy on this dataset and Table 58.2 compares our result with the other state-of-the-art methods.



	Dive	Golf Swing	Kick	Lifting	Ridehorse	Run	Skateboard	Swing-Bench	Swing-Highbar	Walk
Dive	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Golf Swing	0.00	0.83	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.11
Kick	0.00	0.05	0.85	0.00	0.00	0.10	0.00	0.00	0.00	0.00
Lifting	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Ridehorse	0.00	0.08	0.00	0.00	0.75	0.17	0.00	0.00	0.00	0.00
Run	0.00	0.00	0.08	0.00	0.08	0.77	0.08	0.00	0.00	0.00
Skateboard	0.00	0.08	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.08
Swing-Bench	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.90	0.00	0.00
Swing-Highbar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Walk	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.91

Fig. 58.5 Confusion table of our method on UCF Sports dataset

Table 58.2 Recognition results of different methods on the YouTube dataset

Method	Accuracy (%)
Bregonzio et al. [23]	64.0
Liu et al. [19]	71.20
Chakraborty et al. [4]	86.98
Ours	**88.65**

Conclusion

In this paper, a novel method has been proposed to classify human actions in complex scenes. We propose to select interest points by using salient regions. The selected interest points are beneficial to sequential classification because they are inside or around the actors. The proposed method has been validated on two challenging datasets, and the experimental results clearly demonstrate the superiority of our method over previous methods in human action recognition.

References

1. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
2. Zhang Z, Wang C, Xiao B, Zhou W, Liu S (2012) Contextual Fisher kernels for human action recognition. In: International conference on pattern recognition, pp 437–440
3. Raptis M, Soatto S (2010) Tracklet descriptors for action modeling and video analysis. In: European conference on computer vision, pp 577–590
4. Chakraborty B, Holte M, Moeslund T, Gonzàlez J (2012) Selective spatio-temporal interest points. Comput Vis Image Underst 116(3):396–410
5. Liu J, Kuipers B, Savarese S (2011) Recognizing human actions by attributes. In: IEEE conference on computer vision and pattern recognition, pp 3337–3344
6. Zhang Z, Wang C, Xiao B, Zhou W, Liu S (2013) Attribute regularization based human action recognition. IEEE Trans Inf Forensics Secur 8(10):1600–1609
7. Jiang Z, Lin Z, Davis L (2012) A tree-based approach to integrated action localization, recognition and segmentation. In: Trends and topics in computer vision, pp 114–127
8. Kim T, Wong K, Cipolla R (2007) Tensor canonical correlation analysis for action classification. In: IEEE conference on computer vision and pattern recognition, pp 1–8
9. Lin Z, Jiang Z, Davis L (2009) Recognizing actions by shape-motion prototype trees. In: IEEE international conference on computer vision, pp 444–451
10. Cao L, Liu Z, Huang T (2010) Cross-dataset action detection. In: IEEE conference on computer vision and pattern recognition, pp 1998–2005
11. Duchenne O, Laptev I, Sivic J, Bach F, Ponce J (2009) In: IEEE international conference on computer vision, pp 1491–1498
12. Yu T, Kim T, Cipolla R (2010) Real-time action recognition by spatiotemporal semantic and structural forests. In: British machine vision conference
13. Zhang Z, Wang C, Xiao B, Zhou W, Liu S (2012) Action recognition using context-constrained linear coding. IEEE Signal Process Lett 19(7):439–442
14. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2):107–123
15. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE conference on computer vision and pattern recognition, pp 1–8
16. Cheng M, Zhang G, Mitra N, Huang X, Hu S (2011) Global contrast based salient region detection. In: IEEE conference on computer vision and pattern recognition, pp 409–416
17. Felzenszwalb P, Huttenlocher D (2004) Efficient graph-based image segmentation. Int J Comput Vis 59(2):167–181
18. Sullivan M, Shah M (2008) Action MACH: maximum average correlation height filter for action recognition. In: IEEE conference on computer vision and pattern recognition
19. Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos in the wild. In: IEEE conference on computer vision and pattern recognition, pp 1996–2003
20. Wang H, Ullah M, Klaser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. In: British machine vision conference
21. Kovashka A, Grauman K (2010) Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: IEEE conference on computer vision and pattern recognition, pp 2046–2053
22. Le Q, Zou W, Yeung S, Ng A (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE conference on computer vision and pattern recognition, pp 3361–3368
23. Bregonzio M, Li J, Gong S, Xiang T (2010) Discriminative topics modelling for action feature selection and recognition. In: British machine vision conference, pp 1–11
24. Zhang Z, Wang C, Xiao B, Zhou W, Liu S, Shi C (2013) Cross-view action recognition via a continuous virtual path. In: IEEE conference on computer vision and pattern recognition, pp 2690–2697