International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 8, No. 2 (2015), pp. 235-246
http://dx.doi.org/10.14257/ijsip.2015.8.2.23

# Human Action Recognition Based on Global Gist Feature and Local Patch Coding

Yangyang Wang[1], Yibo Li[2] and Xiaofei Ji[2]

*1 College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, People's Republic of China*
*2 College of Automation, Shenyang Aerospace University, Shenyang 110136, People's Republic of China*
*wyy2004101@163.com*

***Abstract***

*Human action recognition has been a widely studied topic in the field of computer. However challenging problems exist for both local and global methods to classify human actions. Local methods usually ignore the structure information among local descriptors. Global methods generally have difficulties in occlusion and background clutter. To solve these problems, a novel combination representation called global Gist feature and local patch coding is proposed. Firstly, Gist feature captures spectrum information of actions in a global view, with spatial relationship among body parts. Secondly, Gist feature located in different grids of the action-centric region is divided into four patches according to the frequencies of action variance. Afterwards on the basis of traditional bag-of-words (BoW) model, a novel formation of local patch coding is adopted. Each patch is encoded independently and finally all the visual words are concatenated to represent high variability of human actions. By combining local patch coding, the proposed method not only solves the problem that global descriptors can not reliably identified actions in complex backgrounds, but also reduces the redundant features in a video. Experimental results performed on KTH and UCF sports dataset demonstrate that the proposed representation is effective for human action recognition.*

***Keywords:*** *action recognition, Gist feature, local patch coding, bag-of-words*

## 1. Introduction

Human action recognition based video is an active research in compute vision and pattern recognition. It has many applications such as public safety monitoring, content-based video indexing and virtual reality technology. However, it is still regarded as a challenging task in realistic scenarios due to viewpoint changes, occlusion, individual variations of human appearance and actions. In order to solve these problems, a large number of techniques and methods based on feature representation are proposed [1-3]. These methods can be partition into two parts: local descriptors with BoW model and global descriptors with template matching.

### (1) Local Descriptors with Bag-of-words Model

Local descriptors describe a set of unordered local feature from a video or image sequences [4]. One of the most known and used features is spatio-temporal interest points (STIP) [5]. On the basis of STIP, local motion or static properties are described, for example, 3-dimensional SIFT descriptor (3DSIFT) [6], histogram of interest point locations (HIPLs) [7] and spatio-temporal descriptor based on histograms of three-dimensional gradients (HOG3D) [8]. If a human action is directly depicted by the collection of all the local descriptors, the dimension of the feature vector would be higher. And the description is too delicate. Therefore, local descriptors with BoW model are

ISSN: 2005-4254 IJSIP

widely adopted and achieve good results [9-10]. The BoW model encodes local descriptors to an unordered visual codebook. Finally actions are modeled as histogram of visual words. These methods based on BoW are robust against occlusion, cluttered background, and minor changes in viewpoint. And they can be extracted without background subtraction or target tracking [11]. However, these local descriptors do not well describe spatial structure of global body actions, because BoW method ignores structural dependencies between body parts [12].

**(2) Global Descriptors with Template Matching**

In contrast to local descriptors, global-based descriptors encode more rich spatial or temporal information within a video sequence. They directly extract and describe the whole properties of human silhouettes or contours. These descriptors are usually classified using template matching. They are also proved to be effective for action recognition [13-14]. A typical global descriptor with template matching is presented by Bobick, *et al*., [15]. They proposed motion history images (MHI) and motion history volume (MHV) for matching. Meng, *et al*., [16] propose a hierarchical MHI. However, this kind of method is sensitive to geometric variations.

(3) **Our Contribution**

Recently some researches have combined the methods based on local and global descriptors to avoid their respectively disadvantages. In [17] holistic silhouettes are used for human activity representation with kernel-induced subspace analysis. Wang, *et al*., [18] modify the HCRF model to combine local patches and large-scale global optical flow features. Inspired by the success of these examples, a novel global Gist feature representation combined local patch coding is proposed.

Gist feature is chosen to the proposed method for three significant reasons: Firstly, Gist feature captures global structure information by filtering an image with different orientations and scales. In the case of realistic scenarios, it can be extracted more reliably than silhouettes feature proposed in [17]. Secondly, the computational time of Gist feature is much less than optical flow features used in [18]. Thirdly, Gist feature can be represented as the concatenation of several local grids with implicit location information. Therefore, combined with Gist feature, a novel framework of local patch code is built on the basis of the traditional BoW model. The Gist features distributed in different locations are respectively quantized into independent local visual words, and then all of the words are concatenated to form the final descriptors. Compared to subspace analysis [17] and HCRF model [18], the global spatial structure of our descriptors can be better kept in a low dimensional feature space.

The main contributions of the proposed method are summarized as follows:

1) The novel representation not only keeps global properties of human action but also shows better tolerance to partial occlusion, viewpoint and noise.

2) By adapting local patch coding scheme for global Gist feature, the characteristic of the patches with salient action variance is strengthened. The main discriminative information is extracted from high dimensional Gist feature. And the redundant information is reduced.

The rest of the paper is organized as follows: the framework is introduced in Section 2. Action representation and recognition is presented in Section 3. The experimental results and analyses are shown in Section 4. Conclusion is drawn in Section 5.

## 2. The Framework of Proposed Method

A graphical overview of our approach is shown in Figure 1. The main processes are of four steps:

 

(1) **Action-centric Region Extraction and Normalization:** Only the area where human located is extracted from each frame of a video. All these areas are respectively normalized to a new rectangular region with fixed size, which is called as action-centric region.

(2) **Global Gist Feature Computation:** A Gist feature is computed to each action-centric region. The global long feature is composed of $m \times n$ Gist vectors located in non-overlapping local grids, where $m \times n$ is the number of the grids in an action-centric region. As shown in Figure 2, the spatial relationships of the vectors are implied in Gist feature.

(3) **Local Patch Coding:** The Gist vectors are partitioned into four patches according to their location distribution. Each patch is respectively encoded and transformed to the form of visual words. Finally a human action is formulated as the concatenation of all of the visual words. By breaking the Gist feature into patches and encoding, the proposed representation becomes highly compact without losing much of the global structure information. Furthermore, in Section 4, traditional BoW method is also used to compare and evaluate the performance of our local coding.

(4) **Action Recognition:** support vector machine (SVM) with radius basis function (RBF) kernel is adopted to recognize the actions. When an unknown video is given, the probabilities of the test video for each trained model are computed. Finally, the unknown video is labeled by using the maximum likelihood model.

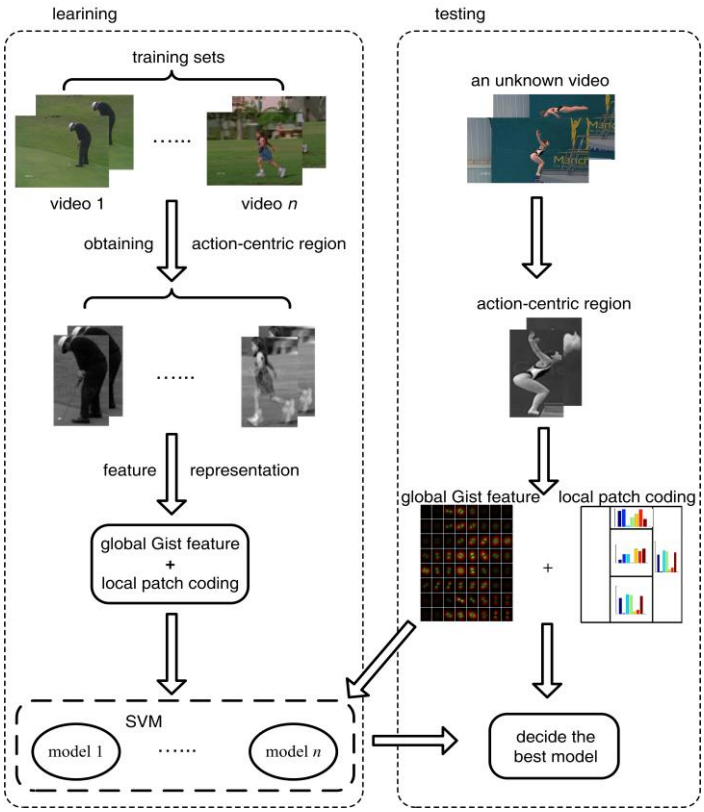


**Figure 1. The Framework of Our Proposed Method**

## 3. Action Recognition with Global Gist Feature and Local Patch Coding

### 3.1. Action-centric Region Extraction and Normalization

By using background substraction, human action region in each frame of a video is separated from background. Some datasets such as UCF sports dataset even

directly provide these regions for researchers to use. In our proposed method, all these regions are respectively normalized to a new 100×60 rectangle. The normalized region is called as action-centric region.

### 3.2. Gist Feature Computation

Gist feature is a kind of filterbank features. It is firstly used in scene classification [19]. And it has also been proven to be effective in objective recognition [20]. Inspired by these works, Gist feature is adopted into the video domain for action recognition. By computing Gist feature, an action-centric region can be represented as the concatenation of spectrum information located in different grids. And these grids in the action-centric region are location dependent. The process consists of the two following two steps:

(1) Gabor filter transfer functions with different orientations and spatial resolution are adopted to filter an action-centric region. The Gabor function is defined as:

$$G(x,y)=\exp(\frac{-(x_1^2+y_1^2)}{2\sigma^{2(s-1)}})\cos(2\pi(F_x x_1+F_y y_1)),$$

where $x_1=x\cos\theta_s+y\sin\theta_s$ , $y_1=-x\sin\theta_s+y\cos\theta_s$ , $(F_x,\ F_y)$ is the frequency of the sinusoidal component, $\sigma$ is standard deviation of Gauss function, $s$ is the number of scales, $\theta_s$ is orientations of the scale $s$. The eight different orientations $(0^\circ,22^\circ,45^\circ,67^\circ,90^\circ,113^\circ,135^\circ,158^\circ)$ are adopted in our proposed method.

(2) The action-centric region is divided into $m\times n$ grids. The average filter response is calculated from each grid, which is called as Gist vector. The dimension of a Gist vector is $s\times\theta$. Finally the whole region can be described by $m\times n$ Gist vectors. These vectors capture the action structure, and also describe the location relationship of local grids.

In this paper, the numbers of orientations and scales are respectively 8 and 4, and the action-centric region is divided into 8×8 grids. Thus the total feature dimensions of an action-centric region are (8×8)×(4×8)=2048. The computational result of Gist feature is illustrated in Figure 2.

The computational time of Gist feature of an action-centric region is about 0.3 second, while the time of global optical flow [18] is about 0.4 second with Intel(R) Xeon(R) @2.4 GHz and MATLAB code. Therefore, compared to the feature used by Wang *et al*. [18], ours are more efficient.

### 3.3. Local Patch Coding

The Gist feature of an action-centric region consists of 64 Gist vectors, and each vector has 32 dimensions. Although the feature has been proved the excellent ability to extract the main characteristic of an objection [20], high-dimensional data easily leads to a lot of redundant information and huge computation burden. Therefore a dimension-reduction method is needed. Considering the generating process of Gist feature, in this paper, an effective local patch coding scheme is presented to create a discriminative and yet compact descriptor.

Each Gist vector corresponds to a particular grid of an action-centric region. Because all the action-centric regions have been normalized, the effect of grids location on action recognition is different. Therefore, according to the structure information and the variation frequency of human body parts, the Gist vectors are divided into different patches. Each patch constructs its special codebook. Finally, all the visual words of the patches are accumulated into final action descriptor. The proposed local patch coding extracts the main discriminative characteristic from high-dimensional global Gist features. And the global spatial structure of human action is also kept. Three steps are employed for local patch coding.

 

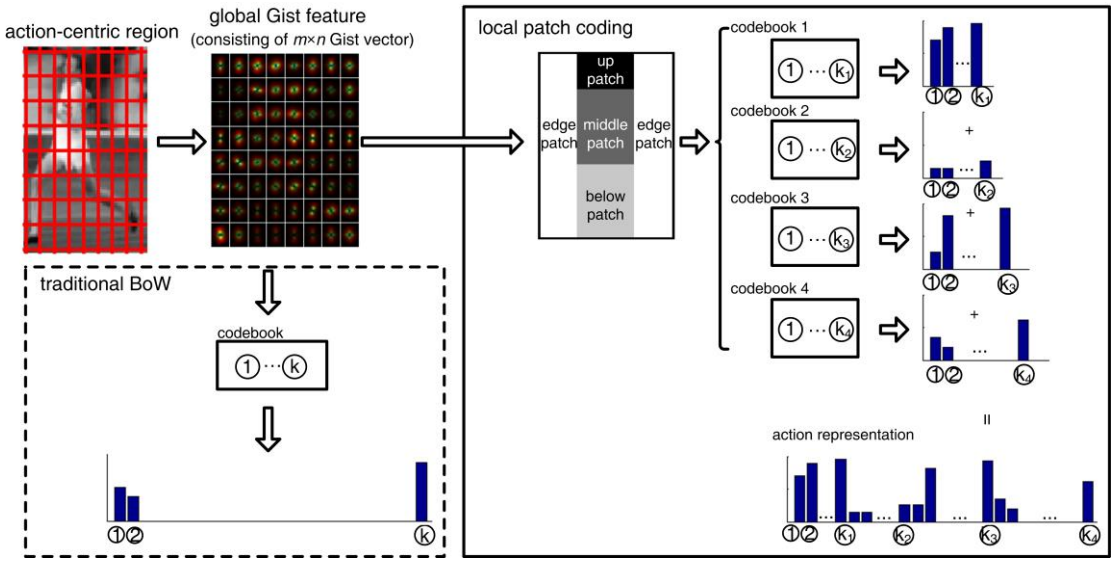


**Figure 2. Description of Global Gist Feature Computation and Local Patch Coding**

(1) **Patch Segmentation:** Each action-centric region is divided into four patches. As shown in Figure 2, the four patches are displayed by different colors. These patches are called as edge patch, up patch, middle patch and below patch respectively.

(2) **Patch-based Coding:** The Gist vectors distributed in each patch alone makes use of bag-of-words method. The codebook size is also different. Usually the variance of human body is more frequent and complex in middle and below patches than in edge and up patches. Therefore, for middle and below patch, larger visual codebooks are used to extract main discriminative Gist representation. And in the other two patches the smaller codebook is applied to get rid of the redundancy to a maximum extent.

(3) **Code Concatenation:** According to four different codebooks, the Gist vectors located in four patches are respectively represented by four groups of visual words. All the words are concatenated together to construct the final action descriptor.

Compared with traditional BoW model, the proposed method can better strengthen local salient action variance. For example, running and walking, both of their variances are concentrated on arms and legs, there exists little change in edge patch. While in the other patches the motion information is rich, with much bigger codebook, the local characteristics of action can be represented more distinctive. While traditional BoW model quantities all the Gist vectors extracted from the action-centric region to words using one codebook, the influence of different local patches on action recognition is not considered. Besides, implicit location relationship is considered to our proposed final descriptor using local patch coding. It can effectively discriminate action variation in a limit feature space.

### 3.4. Action Recognition

In the recognition stage, through the learnt one-*vs*-all SVMs, the test video is assigned the label with maximum votes. RBF kernel is adopted, which can be presented as:

$$K(x,x') = \exp(\frac{\|x-x'\|_2^2}{2\lambda^2}),$$

where $\| \|_2^2$ is the squared Euclidean distance between the two feature vectors x and $x'$. LibSVM toolbox [21] is adopted, and the optimized values of the parameters of SVM models are given by cross validation.

## 4. Experiments and Results

### 4.1. Experimental Settings

Two challenging pubic action datasets are used for evaluating our proposed method: UCF sports dataset [22] and KTH dataset [23]. Action recognition is performed using the leave-one-out cross-validation framework. For UCF sports dataset, followed the experimental settings from Wang, *et al*., [24], the dataset is extended by adding a horizontally flipped version of each video with the purpose of increasing the amount of data samples. And then each original video is tested while training on all other videos together with their flipped versions. Besides, the flipped version of the tested video is removed from the training set. For KTH dataset, testing for this dataset is also proceeded in a leave-one-out framework. Each human is tested while all other actors from all scenarios are trained.

Action region extraction is not our main concern in this work. By using temporal sliding window method, average 30 action regions in each video are selected. For UCF sports dataset these regions are directly taken form the dataset. For KTH dataset, these regions are obtained by using background substraction algorithm in [25]. All of these regions in the two dataset are centered and normalized into 100×60 pixels.

For each 100×60 action-centric region, the global Gist feature is extracted with the 32-dimentional vectors of 8×8 grids. In the process of local patch coding, four patches are divided as shown in Figure 2. Their respective codebook sizes are [$k_1$, $k_2$, $k_3$, $k_4$]. $K = \sum_{i=1}^{4} k_i$, where $K$ is called concatenated codebook size.

### 4.2. Results and Analysis on UCF Sports Dataset

UCF sports dataset consists of 150 broadcast sports videos with a wide range of scenes and viewpoints in unconstrained environments. The videos contain ten different human actions: diving, golfing, kicking, lifting, riding, running, skating, swinging 1(gymnastics, on the pommel horse and floor), swinging 2(gymnastics, on the high and uneven bars) and walking. The recognition task is challenging because of a wide range of scenes and viewpoints.

Because our final descriptor is constructed according to four local codebooks, each codebook size $k_i$ ($i$=1~4) and concatenated codebook size $K$ all have effect on recognition accuracy. The recognition performances with different settings of the [$k_1$, $k_2$, $k_3$, $k_4$] are empirically tested with SVM classifier in two experiments: changing the value of $K$ with $k_i = \frac{K}{4}$ (that is average allocation of local codebooks); allocating different ratios of [$k_1$, $k_2$, $k_3$, $k_4$] under constant $K$. Furthermore, our method is also compared with traditional BoW model to evaluate the performance of the proposed representation.

(1) Action recognition with average allocation of local codebooks

The performance of different $K$ with $k_i = \frac{K}{4}$ is presented in Table 1. In the range of 600 to 1000, our proposed method is not very sensitive to concatenated codebook size $K$. And after exceeding a threshold, the recognition accuracy generally decreases as the $K$ increases.

**Table 1. Recognition Accuracy with Average Allocation of Local Codebooks**

 

| $K=\frac{1}{4}[K,K,K,K]$ | Recognition accuracy (%) |
|---|---|
| 600=[150,150,150,150] | 80.67 |
| 800=[200,200,200,200] | 82 |
| 1000=[250,250,250,250] | 81.33 |
| 1200=[300,300,300,300] | 78.67 |
| 1400=[350,350,350,350] | 79.33 |
| 1600=[400,400,400,400] | 76.67 |

(2) Action recognition with different ratios of local codebooks

By analyzing the characteristics of action types in UCF sports dataset, motion variance happens more frequently in middle patch and below patch than the other two patches. Therefore, the bigger size of codebooks is chose for middle patch and below patch.

It can be drawn from Table 1 that the descriptive capability is preserved when the value of $K$ is in the range of 600 to 1000. Therefore, taken $K$=800 for example, the recognition results with different ratios of [$k_1$, $k_2$, $k_3$, $k_4$] are shown in Table 2. By experimental verification, the accuracy is usually higher when the ratio of four local codebooks is 1:1:2:4. The best overall mean accuracy is 86% for UCF sports dataset with [100,100,200,400]. Its recognition accuracy is obtained by computing the percentage of the numbers of action which are rightly recognition to all the 150 testing actions. The confusion matrix is depicted in Figure 3.

(3) Comparison with traditional BoW model

To the same global Gist features, Figure 4 reports a comparison of recognition accuracy between our proposed method and traditional BoW model. The same k-means algorithm is adopted to construct visual codebook. In traditional BoW model, all of the Gist vectors are clustered into one codebook. In order to keep the consistency of codebook size, the concatenated codebook size $K$ is set with average allocation of local codebooks in our proposed method. Bold line indicates the accuracy of traditional BoW model, and the dash line represents the accuracy of our proposed method. The experiments prove that our method performed better than traditional BoW model by considering different variance frequencies of actions in different local patch and the spatial structure information among Gist vectors.

**Table 2. Recognition Accuracy with Different Ratios of Local Codebooks**

| [$k_1$, $k_2$, $k_3$, $k_4$] | Recognition accuracy (%) |
|---|---|
| [50,100,200,450] | 80.67 |
| [100,50,200,450] | 78.67 |
| [100,100,200,400] | 86 |
| [100,100,400,200] | 81.33 |
| [100,100,300,300] | 83.33 |
| [100,150,250,300] | 80 |
| [100,150,300,250] | 82 |
| [150,100,250,300] | 82.67 |
| [150,100,300,250] | 81.33 |

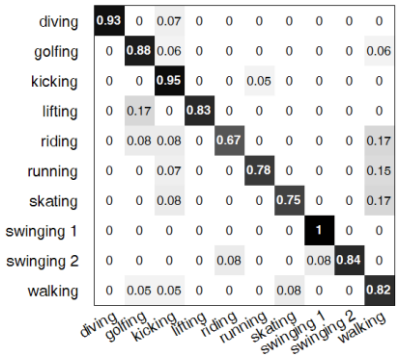


**Figure 3. Confusion Matrix for UCF Sports Dataset**

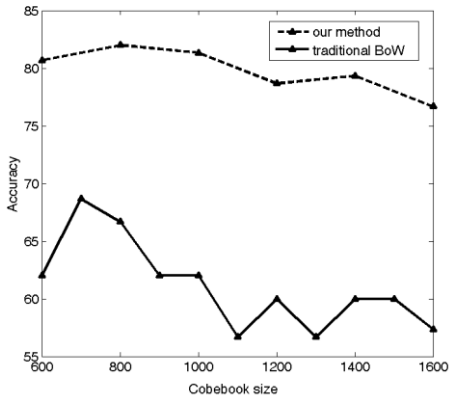


**Figure 4. Recognition Accuracy using Our Proposed Method and Traditional BoW Model**

(4) Comparison with other related methods

Table 3 shows a comparison of our proposed method with other related methods for UCF sports dataset in recent years. The methods in Table 3 are all based on bag-of-features or global features. Our recognition result outperforms other methods by combing global Gist feature and local patch coding.

**Table 3 Comparison of Our Method with Related Methods for UCF Sports Dataset**

| Author | Method | Accuracy (%) |
|---|---|---|
| Raptis *et al*. [26] | Dense trajectories + Clusters | 79.4 |
| Yu *et al*. [27] | WSAP trajectories + Concatenated BoF | 81.07 |
| Ullah *et al*. [28] | Local motion descriptor + BoW | 82.5 |
| Lan *et al*. [29] | Figure-centric visual word | 79.1 |
| Our proposed method | Global Gist feature + Local patch coding | 86 |

### 4.3. Results and Analysis on KTH Dataset

KTH dataset contains six types of human action: walking, jogging, running, boxing, handwaving and handclapping. Each action is performed by 25 persons in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors.

The best overall mean accuracy is 91.83% for KTH dataset with $K$=[50, 200, 300, 350]. Its corresponding confusion matrix is depicted in Figure 5(a). Compared 86% for UCF sports dataset with $K$= [100,100,200,400]. The number of $k_1$ in KTH is

 

smaller than UCF sports dataset. The reason is that the actions of UCF are collected from various sports videos. The motion variances are more diversiform, taken diving and swing actions for example, more motion variance of body parts are appeared in edge patch. Wile for the actions such as walking, motion variance are more centered in middle and below patches.

Besides, the codebook size $K$=900 which obtains the best results of our proposed method is also used for traditional BoW model. Its recognition accuracy is 73.3%. Figure 5(b) shows the confusion matrix of traditional BoW model. The results demonstrate that our proposed local patch coding provides better performance than traditional BoW model.

Table 4 summarizes action recognition accuracies using related methods for KTH dataset. All these methods remain in the frameworks of the BoW or codebook model. Our recognition accuracy achieves the better result than most other methods. Yu's result on KTH dataset is higher than ours while the recognition accuracy of UCF sports dataset is lower by about 5%. In general our proposed method is easy to implement and the performance for action recognition is satisfactory.


**Figure 5. Confusion Matrix for KTH Dataset**

**Table 4 Comparison of Our Method with Related Methods for KTH Dataset**

| Author | Method | Accuracy (%) |
|---|---|---|
| Yu *et al*. [27] | WSAP trajectories + Concatenated BoF | 96.3 |
| Oikonomopoulos *et al*. [30] | Spatiotemporal shape model + Class-specific codebook | 88 |
| Ye *et al*. [31] | Local feature group + BoW with discriminate group distance | 91.6 |
| Cheng *et al*. [32] | Temporal relations + Bag of groups | 89.7 |
| Our proposed method | Global Gist feature + Local patch coding | 91.83 |

## 5. Conclusion

This paper proposed a novel action representation by combining global Gist feature and local patch coding based on BoW model. High-dimensional global feature are transformed into the form of ordered and compact concatenated visual words through patch segmentation and local coding. The advantage is that not only main global properties of human action are kept but also the influence of occlusion and noise is reduced. Evaluations on two challenging realistic scenarios action datasets, KTH dataset and UCF sports dataset, prove that our proposed method has the capability of recognizing diverse actions in a large variety of scenarios.

## Acknowledgements

The project supported by the National Natural Science Foundation of China (No. 61103123), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and the Scientific Research General Project of Education Department of Liaoning Province, China (No. L2014066).

## References

[1]. Y. Li, T. Sun and X. Jiang, "Human action recognition based on oriented gradient histogram of slide blocks on spatio-temporal silhouette", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 5, no. 3, **(2012)**.
[2]. D. Weinland, R. Ronfard and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition", Computer Vision and Image Understanding, vol. 115, no. 2, **(2011)**.
[3]. X. Ji and H. Liu, "Advances in view-invariant human motion analysis: a review, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, no. 1, **(2010)**.
[4]. A. Perina, M. Cristani and V. Murino, "Learning natural scene categories by selective multi-scale feature extraction. Image and Vision Computing, vol. 28, no. 6, **(2010)**.
[5]. G. Willems, T. Tuytelaars and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector", Proceedings of the 10th European Conference on Computer Vision, **(2008)** October 12-18, Marseille, France.
[6]. P. Scovanner, S. Ali and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition", Proceedings of the 15th International Conference on Multimedia, **(2007)** September 24-29, Augsburg, Germany.
[7]. X. Yan and Y. Luo, "Recognizing human actions using a new descriptor based on spatial-temporal interest points and weighted-output classifier", Neurocomputing, vol. 87, **(2012)**.
[8]. P. Bilinski and F. Bremond Evaluation of local descriptors for action recognition in videos, Proceedings of the 8th International Conference on Computer Vision Systems, **(2011)** September 20-22, Sophia Antipolis, France.
[9]. X. Zhen and L. Shao, "A local descriptor based on Laplacian pyramid coding for action recognition", Pattern Recognition Letters, vol. 34, no. 15, **(2013)**.
[10]. V. Bettadapura, G. Schindler, T. Plötz and I. Essa, "Augmenting bag-of-words: data-driven discovery of temporal and structural information for activity recognition," IEEE Conference on Computer Vision and Pattern Recognition, **(2013)** June 25-27, Portland, Oregon, USA.
[11]. J. Odobez, R. Emonet and R. Tavenard, "Time-Sensitive topic models for action recognition in videos", Proceedings of the 20th IEEE International Conference on Image Processing, September 15-18, Melbourne, Australia.
[12]. A. Yao, J. Gall and L. Van Gool, "A hough transform-based voting framework for action recognition", IEEE Conference on Computer Vision and Pattern Recognition, **(2010)** June 13-18, San Francisco, California, USA.
[13]. J. Yang and Z. Ma, "Action recognition using local spatio-temporal oriented energy features and additive kernel SVMs", International Journal of Electronics and Electrical Engineering, vol. 2, no. 2, **(2014)**.
[14]. G. Goudelis, K. Karpouzis and S. Kollias, "Exploring trace transform for robust human action recognition", Pattern Recognition, vol. 46, no. 12 **(2013).**
[15]. A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, **(2001).**
[16]. H. Meng, N. Pears and C. Bailey, "A human action recognition system for embedded computer vision application", IEEE Conference on Computer Vision and Pattern Recognition, **(2007)** June 18-23, Minneapolis, Minnesota, USA.
[17]. L. Wang and D. Suter, "Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model", IEEE Conference on Computer Vision and Pattern Recognition, **(2007)** June 18-23, Minneapolis, Minnesota, USA.
[18]. Y. Wang and G. Mori, "Hidden part models for human action recognition: probabilistic versus max margin", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 7, **(2011).**
[19]. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope", International Journal of Computer Vision, vol. 42, **(2001).**
[20]. A. Torralba, K. P. Murphy, W. T. Freeman and M. A. Rubin, "Context-based vision system for place and object recognition", IEEE International Conference on Computer Vision, **(2003)** October 14-17, Nice, France.
[21]. C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, vol. 2, **(2011).**
[22]. M. D. Rodriguez, J. Ahmed and M. Shah, "Action mach: a spatiotemporal maximum average correlation height filter for action recognition", IEEE Conference on Computer Vision and Pattern Recognition, **(2008)** June 24-26, Anchorage, Alaska, USA.

 

[23]. C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach", Proceedings of the 17th International Conference Pattern Recognition, **(2004)** August 23-26, Cambridge, England, UK.
[24]. H. Wang, M. M. Ullah, A. Klaser, I. Laptev and C. Schmid, "Evaluation of local spatio-temporal features for action recognition", British Machine Vision Conference, **(2009)** September 7-10, London, UK.
[25]. M. Bregonzio, S. Gong and T. Xiang, "Recognising action as clouds of space-time interest points", IEEE Conference on Computer Vision and Pattern Recognition, **(2009)** June 20-26, Florida, USA.
[26]. M. Raptis, I. Kokkinos and S. Soatto, "Discovering discriminative action parts from mid-level video representations", IEEE Conference on Computer Vision and Pattern Recognition, **(2012)** June 16-21; Providence, Rhode Island, USA
[27]. J. Yu, M. Jeon and W. Pedrycz. Weighted feature trajectories and concatenated bag-of-features for action recognition. Neurocomputing, vol. 131, **(2014).**
[28]. M. M. Ullah and I. Laptev. Actlets: a novel local representation for human action recognition in video. IEEE International Conference on Image Processing, **(2012)** September 30 - October 3, Florida, USA
[29]. T. Lan, Y. Wang and G. Mori. Discriminative figure-centric models for joint action localization and recognition. IEEE International Conference on Computer Vision, **(2011)** November 6-13, Barcelona, Spain
[30]. A. Oikonomopoulos, I. Patras and M. Pantic, "Spatiotemporal localization and categorization of human actions in unsegmented image sequences", IEEE Transactions Image Processing, vol. 20, no. 4, **(2011).**
[31]. Y. Ye, L. Qin, Z. Cheng and Q. Huang, "Recognizing realistic action using contextual feature group", The Era of Interactive Media, **(2013).**
[32]. G. Cheng, Y. Wan, W. Santiteerakul, S. Tang and B. P. Buckles, "Action recognition with temporal relationships", IEEE Conference on Computer Vision and Pattern Recognition Workshop, **(2013)** June 23-24, 28, Portland, Oregon, USA

## Authors

**Yangyang Wang** received her M.S. degree from the Shenyang Institute of Aeronautical Engineering, in 2006. She is currently a graduate student studying for Ph.D. degree in the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics. She has published over 10 technical research papers. Her research interests include vision analysis and pattern recognition.

**Yibo Li** received his M.S. and Ph.D. degrees from the Nanjing University of Aeronautics and Astronautics and Northeastern University in 1986 and 2003, respectively. Since 1999, he holds the position of Full Professor at Shenyang Aerospace University. He has published over 100 technical research papers and books. More than 30 research papers have been indexed by SCI/EI. His research interests include biometrics recognition, image processing, and fight control systems.

**Xiaofei Ji** received her M.S. and Ph.D. degrees from the Liaoning Shihua University and University of Portsmouth in 2003 and 2010, respectively. From 2013, she holds the position of Associate Professor at Shenyang Aerospace University. She is the IEEE member, has published over 40 technical research papers and 1 book. More than 20 research papers have been indexed by SCI/EI. Her research interests include vision analysis and pattern recognition.