International Journal of Automation and Computing

11(5), October 2014, 500-509
DOI: 10.1007/s11633-014-0831-4

Study of Human Action Recognition Based on Improved Spatio-temporal Features

Xiao-Fei Ji[1] Qian-Qian Wu[1] Zhao-Jie Ju[2] Yang-Yang Wang[1]

[1]School of Automation, Shenyang Aerospace University, Shenyang 110136, China
[2]School of Computing, University of Portsmouth, Portsmouth PO54BP, UK

Abstract: Most of the exist action recognition methods mainly utilize spatio-temporal descriptors of single interest point while ignoring their potential integral information, such as spatial distribution information. By combining local spatio-temporal feature and global positional distribution information (PDI) of interest points, a novel motion descriptor is proposed in this paper. The proposed method detects interest points by using an improved interest point detection method. Then, 3-dimensional scale-invariant feature transform (3D SIFT) descriptors are extracted for every interest point. In order to obtain a compact description and efficient computation, the principal component analysis (PCA) method is utilized twice on the 3D SIFT descriptors of single frame and multiple frames. Simultaneously, the PDI of the interest points are computed and combined with the above features. The combined features are quantified and selected and finally tested by using the support vector machine (SVM) recognition algorithm on the public KTH dataset. The testing results have showed that the recognition rate has been significantly improved and the proposed features can more accurately describe human motion with high adaptability to scenarios.

Keywords: Action recognition, spatio-temporal interest points, 3-dimensional scale-invariant feature transform (3D SIFT), positional distribution information, dimension reduction.

1 Introduction

In recent years, visual based human action recognition has gradually become a very active research topic. Analysis of human actions in videos is considered a very important problem in computer vision because of such applications as human-computer interaction, content-based video retrieval, visual surveillance, analysis of sport events[1]. Due to the complexity of the action, such as different body wearings and habits leading to different observations of the same action, the camera movement in the external environment, illumination change, shadows, viewpoint, these influences of factors make action recognition still a challenging project[2, 3].

The representation of human motion in video sequences is crucial to action recognition. Other than having enough discrimination between different categories, reliable motion features are also required to deal with rotation, scale transform, camera movement, complex background, shade, etc. At present, most commonly used features in action recognition are based on motion, such as optical flow[4, 5], motion trajectory[6−8], or based on appearance shape, such as silhouette contour[9, 10]. The former features are greatly influenced by illumination and shadow. The latter features rely on accurate localization, background subtraction or tracking, and they are more sensitive to noise, partial occlusions and variations in viewpoint. Compared with the former two kinds of features, local spatio-temporal features are somewhat invariant to changes in viewpoint, person appearance and partial occlusions[11]. Due to their advantage, local spatio-temporal features based on interest points are more and more popular in action recognition[12−15].

Spatio-temporal interest points are those points where the local neighborhood has a significant variation in both spatial and temporal domains. Most of local spatio-temporal feature descriptors are the extension of the information extracted from previously 2D space to 3D spatio-temporal based on the interest points detected by Laptev and Lindebery[13] or Dollár et al.[12]. They capture motion variation in space and time dimensions in the neighborhood of the interest points. Up to now, many efforts have been devoted to the description of the spatio-temporal interest points. The most common descriptions are scale-invariant feature transform (SIFT)[16], speeded-up robust features (SURF)[17], which have advantages of scale, affine, view and rotation invariance. Dollár et al.[12] applied the cuboids and selected smooth gradient as a descriptor. Later, Scovanner et al.[18] put forward the 3D SIFT to calculate spatio-temporal gradient direction histograms for each pixel within its neighborhood. Another extension to the SIFT was proposed by Kläser et al.[19], based on a histogram of 3D gradient orientations, where gradients are computed using an integral video representation. Williems et al.[20] extended the SURF descriptor to video, by representing each cell as a vector of weighted sums of uniformly sampled responses to Haar wavelets along the three axes.

In the process of recognition, with the extraction of cuboids feature descriptor, Dollár et al.[12] adopted the principal component analysis (PCA) algorithm to reduce feature dimension and finally made use of the nearest neighbor classifier and support vector machine (SVM) to recognize human actions on KTH dataset. Niebles et al.[14] consid-

Regular paper
Special Issue on Massive Visual Computing
Manuscript received January 14, 2014; accepted June 20, 2014
This work was supported by National Natural Science Foundation of China (No. 61103123) and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

ered videos as spatio-temporal bag-of-words by extracting space-time interest points and clustering the features, and then used a probabilistic latent semantic analysis (PLSA) model to localize and categorize human actions. Li and Du[21] got interest points from Harris detector and then extracted 3D SIFT descriptor. In the recognition process, they made use of SVM with leave-one-out method and also did the experiment on the KTH dataset. The above recognition methods have achieved good recognition results, but most studies only stayed on the previous description of the interest points, mainly utilized local spatio-temporal descriptors of single interest point while ignoring its overall distribution information in the global space and time. The interest points represent the key position of the human body movement. So the distribution of interest points change according to the human motion. And its implication of sport information also changes accordingly. In addition, it cannot simply rely on the local feature of spatio-temporal interest points to represent the target motion when the motion lacks time dimension information. Bregonzio et al.[22] defined a set of features which reflect the interest point distribution based on different temporal scales. In their study, global spatio-temporal distribution of interest points was studied but the excellent performance of local descriptor was also abandoned. When the influence factors interfere body movement, it cannot reliably represent the action by only depending on the global information. These aforementioned experiments were all conducted on the KTH dataset for the recognition test, but the adaptability of feature applied in different scenarios on the KTH was not studied and discussed. Moreover, despite the popularity of [12], it tends to generate spurious detection background area surrounding the object boundary due to the shadow and noise, and then the subsequent recognition process is affected too.

Based on the above discussions, a novel feature is proposed in this paper to represent human motion by combining local and global information, i.e., a combination of 3D SIFT descriptor and the spatio-temporal distribution information based on interest points. First, an improved detecting method[22] is used to detect spatio-temporal interest points, which is different from [12] and can effectively avoid the error detecting in the background. Then, these interest points are represented by 3D SIFT descriptor and the positional distribution information of these interest points is calculated at the same time. In order to achieve a perfect combination, the dimension reduction issue is solved on the descriptor twice, which is based on single frame and multiple frames. Then, the processed descriptor is combined with the positional distribution information of interest points. Finally, the combined features are quantified and selected to obtain more concise feature descriptors. The 3D SIFT descriptor contains human body posture information and motion dynamic information, and it describes the local feature of action both in spatio and temporal dimension. As a result of the feature extracted in the key points of motion, it is not affected by any change in human body shape, motion directions, etc. So the feature has good adaptability and robustness in complex motion scenarios. The positional distribution information of the interest point reflects the motion global information by using various location and ratio relationships of the two areas of human body movement and interest point distribution. Different from previous methods of describing the appearance and shape in space, this paper doesn′t directly pick up the shape information. So when the human appearance and shape change, the location and ratio relationship of the two areas are not directly affected. Therefore, the positional distribution information of interest point proposed by this paper is more adaptive to motion description in space. Finally, the proposed motion descriptor by combining the above mentioned local feature with global information is tested by using an SVM recognition algorithm on the public dataset of KTH. Furthermore, the adaptability of the proposed method is discussed by testing in different and mixed scenarios of KTH dataset. By comparing with the related and similar research works in recent years, the results have verified that the proposed method is better in terms of strong robustness and adaptability.

The rest of the paper is organized as follows. In Section 2, the detection method for spatio-temporal interest points is introduced. Section 3 provides a detailed explanation of 3D SIFT descriptor and positional distribution information of interest points, as well as the process of feature dimension reduction, quantification and selection. Section 4 gives experimental results and analysis. Finally, Section 5 concludes the paper.

2 Interest point detection

In computer vision, interest points represent the location which has severe changes in space and time dimensions and is considered to be salient or descriptive for the action captured in a video. Among various interest point detection methods, the most widely used for action recognition is the one proposed by Dollár et al.[12]. The method calculates function response values based on the combination of Gabor filter and Gaussian filter, and the extreme values of local response can be considered as spatio-temporal interest points in the video.

Dollár′s method is effective to detect the interest points of human motion in video. However, it is prone to false detection due to video shadow and noise, and spurious interest points are easy to occur in the background. It is particularly ineffective to camera movement, or camera zooming. Some of the drawbacks are highlighted in the examples as red square slices shown in Fig. 1 (c).

These drawbacks are due to the shortcomings of its detector, especially the Gabor filtering which does feature extraction only on the time axis while ignoring the dynamic movement in the prospects. To overcome these shortcomings, we utilize a different interest point detector[22] which explores different filters for detecting salient spatio-temporal local areas undergoing complex motion to get a combined filter response. More specifically, our detector facilitates saliency detection and consists of the following three steps:

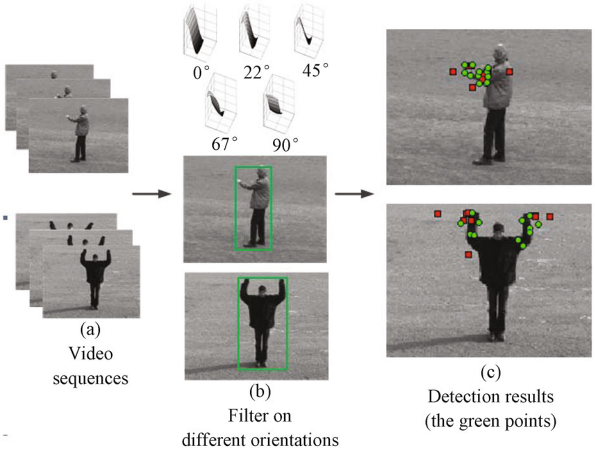

0°
22°
45°
67°
90°
(a)
Video sequences
(b)
Filter on different orientations
(c)
Detection results (the green points)

Fig. 1 Interest point detection process

Step 1. Frame differencing is utilized to detect the regions of interest as shown in Fig. 1 (b).

Step 2. Utilize 2D Gabor filter for generating five different directions (0°, 22°, 45°, 67°, 90°) filter templates. The example is presented in Fig. 1 (b).

Step 3. Filter on the detected regions of interest is got in Step 2. Use 2D Gabor filters with five different orientations obtained by step 2) in both spatial and temporal domains to give a combined filter response.

Fig. 1 (c) shows the examples of our interest point detection results (circle points) and the Dollár′s[12] (square points). It is evident that our detected interest points are much more meaningful and descriptive compared to those detected using the Dollár detector. The color graph of Fig. 2 can be seen in the electronic version.

3 Action representation

3.1 3D SIFT descriptor

For calculating the SIFT descriptor, spatio-temporal cube is extracted from the interest point as the center in video sequences and is divided into fixed-size unit sub cubes. Then, the spatio-temporal gradient histogram of each unit cube is calculated by using faceted sphere. Finally, the 3D SIFT descriptor is formed by combining all the unit cube histograms[18]. In this paper, the 12 × 12 × 12 pixel-size cube is divided into $2\times2\times2$ sub cubes, as shown in Fig. 2.

3-dimensional gradient magnitude at (x, y, t) is computed by

$$M(x,y,t)=\sqrt{L^2{}_x+L^2{}_y+L^2{}_t} \tag{1}$$

where $L_x = L(x+1, y, t) - L(x-1, y, t)$, $L_y = L(x, y+1, t) - L(x, y-1, t)$ and $L_t = L(x, y, t+1) - L(x, y, t-1)$ stand for the gradients in the x, y and t directions, respectively.

According to the amplitude weight in the center $V(v_{xi}, v_{yi}, v_{ti})$ of each sphere face, three maximums are chosen and added to the corresponding direction of the gradient histogram by using

$$\begin{aligned} mag = M(x,y,t)G(x',y',t')V(v_{xi},v_{yi},v_{ti})\\ i=1,2,\cdots,32 \end{aligned} \tag{2}$$

where (v_{xi}, v_{yi}, v_{ti}) is the center coordinate of each surface on the faceted sphere which relies on the current pixel as the center. $G(x', y', t') = \mathrm{e}^{\frac{x'^2+y'^2+t'^2}{2\sigma^2}}$ serves as the gradient weight, x', y' and t' are the difference values between the interest point and the current pixel in the neighborhood.

This paper adopts a 32 faceted sphere and 32 gradient directions for the descriptor, so the feature dimension of each sub cube is 32. The initial whole features of each point are of 256 dimensions.

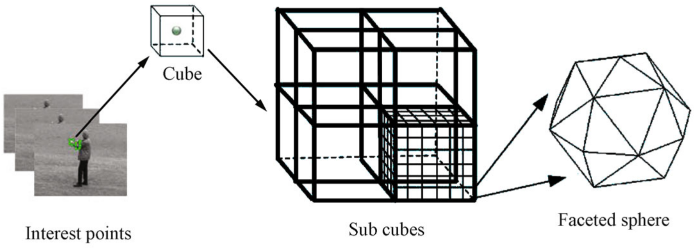


Fig. 2 Description process of 3D SIFT

3.2 Positional distribution information of interest points

In terms of interest points in each frame, the positional distribution information is closely related to the body motion, it also reflects the amplitude range of action, relevance of human body location and motion part region. So the positional distribution information of interest points are extracted as another kind of action information. The specific process is described as below.

As shown in Fig. 3, we select the example actions from KTH dataset, then the distribution region of interest points and body location area are detected in each frame. The region and area are drawn with yellow (Y) and red (R) boxes respectively. Then the related positional distribution information is calculated with these two areas and expressed by $PDI = [D_{\rm ip}, R_{\rm ip}, R_{\rm ren}, Vertic_{\rm dist}, Orizon_{\rm dist}, W_{\rm ratio}, H_{\rm ratio}, Over_{\rm lap}]$. A calculation method of the features is shown in Table 1.

In order to remove the influence of several stray individual points on the overall feature, a stray filtering process is utilized for all the interest points before extracting the above positional feature. The points whose distances to the region centroids are over a certain threshold value are removed to ensure the validity and reliability of the extracted features. The positional distribution information of interest points is extracted from each frame to represent and reflect the whole attributes of the motion.

3.3 Motion features

In Section 3.1, the 3D SIFT descriptor of each interest point is a 256 dimensional vector. If the number of interest points in each frame is N, the dimension of the features is $N \times 256$ to represent the spatio-temporal information in this frame. The dimension of feature is so high that it cannot reasonably combine the distribution information PDI obtained by Section 3.2. Therefore, dimension reduction is performed on this part information, as shown in Fig. 4. Furthermore, in order to remove redundant data and make features more concise, combined features are processed by using quantization and selection. The following five steps are listed:

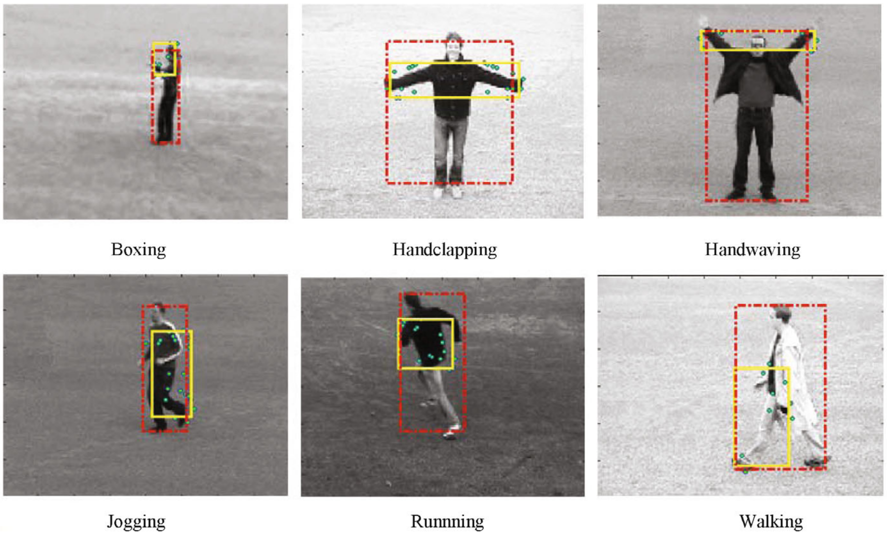


Fig. 3 Distribution of interest points

Table 1 Calculation methods of PDI feature

PDI	Calculation method
$D_{\rm ip}$	The total number of points normalised by the area Y
$R_{\rm ip}$	The height and width ratios of Y
$R_{\rm ren}$	The height and width ratios of R
$Vertic_{\rm dist}$	The vertical distance between the geometrical centre (centroid) of Y and R
$Orizon_{\rm dist}$	The horizontal distance between the geometrical centre (centroid) of Y and R
$W_{\rm ratio}$	The width ratio between the two areas Y and R
$H_{\rm ratio}$	The height ratio between the two areas Y and R
$Over_{\rm lap}$	The ratio by the amount of overlap and total width between Y and R

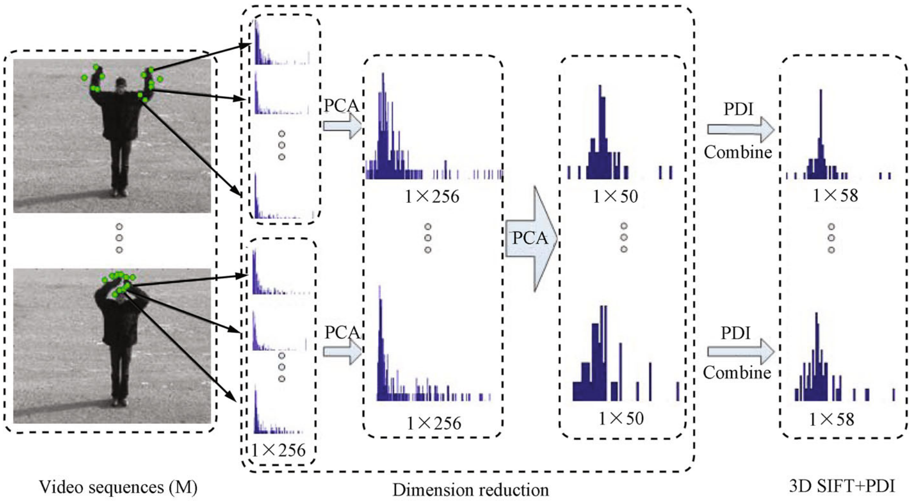


Fig. 4 Description of motion feature

Step 1. Single frame dimension reduction. Principle component analysis is used to perform longitudinal dimension reduction for 3D SIFT descriptor extracted from interest points in the same frame. It means that $N \times 256$ features can be reduced (N is the number of interest points in this frame) to 1×256 by gathering principal components of all the descriptors for each frame. The single frame dimension reduction is helpful to achieve a whole description of the motion in the frame. Although it will lose part of information, the loss of the information is acceptable when N is chosen as 20 in our experiment.

Step 2. Multi-frame dimension reduction. Horizontal dimension reduction is done on the preprocessed descriptors got by Step 1. The dimension reduction is used again on all the frames to set $M \times 256$ (M is the total number of video frames) to $M \times 50$.

Step 3. Feature combination. Make the combination of 1×50 spatio-temporal features and the corresponding positional distribution information (PDI) of interest points in each frame, finally we get 58 dimension features for each frame (3D SIFT+ PDI).

Step 4. Feature quantization. The linear quantization is utilized on $M \times 58$ dimension feature of each video to product a histogram containing n ($n < M$) bins. Thus, it is turned into $n \times 58$ dimension feature.

Step 5. Feature selection. Compute the mean of distances from different people performing the same action and make the arrangement, then select the front location S ($S < n \times 58$) features as the final features.

4 Algorithm verification and result analysis

In this section, experiments are performed on the KTH dataset with the improved spatio-temporal feature. By comparing with the most recent reports associated with the related features and dataset, the outstanding performance of the proposed algorithm is demonstrated in this section.

4.1 Recognition algorithm

SVM[23], as the data classification of statistical learning method, has intuitive geometric interpretation and good generalization ability, so it has gained popularity in visual pattern recognition.

According to the theory, SVM is developed from the theory of structural risk minimization, shown as

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{R}\varepsilon_i$$
$$y_i(w, \phi(x_i) - b) \geqslant 1 - \varepsilon_i,\ 0 \leqslant \varepsilon_i \leqslant 1. \tag{3}$$

For a given sample set x_i, $y_i \in \{-1, 1\}$, $i = 1, \cdots, n$, where $x_i \in \mathbf{R}^N$ is a feature vector and y_i is its class label, $(\phi(x_i), \phi(x_j)) = k(x_i, x_j)$ is the kernel function, k is corresponding to the dot product in the feature space, and transformation ϕ implicitly maps the input vectors into a high-dimensional feature space. Define the hyperplane $(w, \phi(x)) - b = 0$ to make a compromise between class interval and classification errors when the sample is linearly inseparable. Here, ε_i is the i-th slack variable and C is the regularization parameter. This minimization problem can be solved using Lagrange multiplier and Karush-Kuhn-Tucker (KKT) conditions, and the dual function is written as

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(\phi(x_i), \phi(x_j)) \tag{4}$$

$$C \geqslant \alpha_i \geqslant 0, i = 1, \cdots, n, \qquad \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{5}$$

where $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^{\mathrm{T}}$ is a Lagrange multiplier which corresponds to $y_i(w, \phi(x_i) - b) \geqslant 1 - \varepsilon_i, \varepsilon_i \geqslant 0$. We select the kernel function $k(x_i, x_j) = \mathrm{e}^{-\frac{(x_i - x_j)^2}{2\sigma^2}}$ and put it into (4) to get the final decision function as

$$y(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i y_i k(x_i, x) - b). \tag{6}$$

Instead of establishing SVMs between one against the rest types, we adopt the method for one against one to establish an SVM between two categories. The category the current sample belongs to is determined by the decision function, and its final type is decided by the category with the highest vote.

4.2 Dataset

To test our proposed approach for action recognition, we choose the standard KTH dataset, which is one of the most popular benchmark datasets for evaluating action recognition algorithms. This dataset is challenging because there are large variations in human body shape, view angles, scales and appearance. As shown in Fig. 5, the KTH dataset contains six types of different human actions performed by 25 different persons: boxing, hand clapping, hand waving, jogging, running, and walking. And the sequences are recorded in four different scenarios: outdoors (SC1), outdoors with scale variations (SC2), outdoors with different clothes (SC3), and indoors with lighting variations (SC4). There are obvious changes of visual sense or view between different scenarios, and the background is homogeneous and static in most sequences with some slight camera movement. The sequences are downsampled to the spatial resolution of 160 × 120 pixels. The examples of the above four scenarios are shown in Fig. 5. Apparently, due to the change of camera zooming situation, the size of human body changes a lot in SC2 (captured in conditions of zooming out (t_1) and zooming in (t_2)). Furthermore, the person in SC3 presents larger changes in body appearance because of the different dress or with a bag.

4.3 Testing results in portion scenario

In this part, recognition experiments are performed by using combined feature of 3D SIFT and PDI on four scenarios (SC1, SC2, SC3 and SC4) according to KTH dataset. The proposed feature is tested in various scenarios to verify its reliability for motion description and adaptability to scenarios. The leave-one-out cross validation method is adopted throughout the process. Six actions of each actor are used in turn as test samples, and the rest of all the actions are used as the training. The process continues until all the actions are completed tested. The experimental results are shown in Table 2.

The experimental results show that the feature of 3D SIFT has a better discriminative ability than PDI feature. SC1 and SC4 are more stable than the other two scenarios. We obtained the same recognition rate (96%) by using 3D SIFT and combined features (3D SIFT+PDI). In spite of the influences of motion direction changes and indoor lighting, the 3D SIFT (after the process of dimension reduction and quantization) is still a good feature description for motion. It also shows that 3D SIFT has good adaptability and robustness to motion direction, position, speed, etc.

Table 2 Testing results in portion scenarios

Scenario	3D SIFT	PDI	3D SIFT+PDI
SC1	0.9600	0.8000	0.9600
SC2	0.8867	0.8268	0.9200
SC3	0.8542	0.7569	0.9167
SC4	0.9600	0.9000	0.9600

In SC2 and SC3, scenarios become more complex. Not only the human body exists scale variations with camera zooming in SC2, but also there are 45-degree view changes in jogging, running and walking. In SC3, the human body shape changes with different wearings, and the phenomenon of inhomogeneous background even emerges. These above situations make the motion area and position distribution of interest points change obviously as well. So the combined features (3D SIFT+PDI) have certain advantages than 3D SIFT to describe motion in these scenarios. The recognition

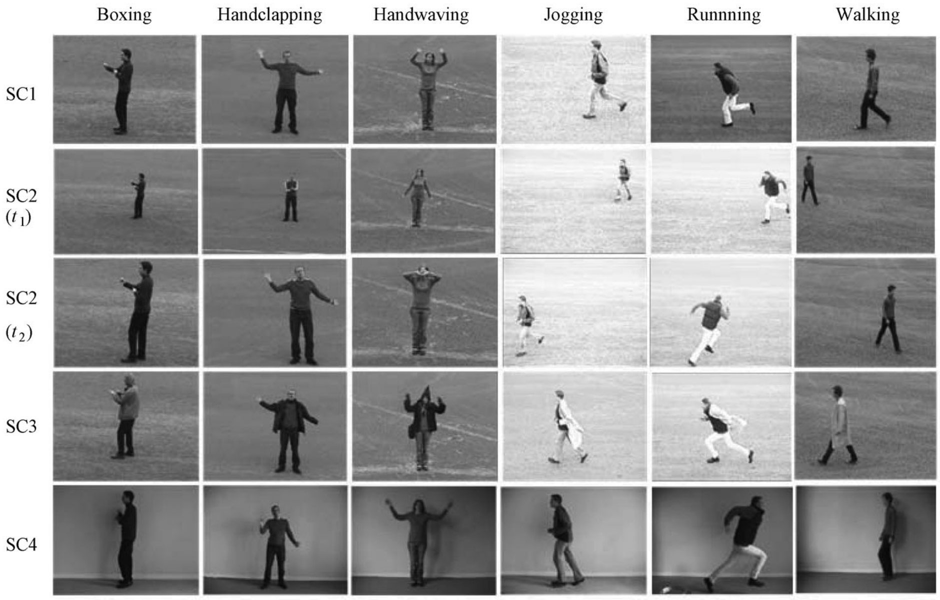


Fig. 5 Description of motion feature

rate is greatly increased by using combined features (3D SIFT+PDI) compared to 3D SIFT. The recognition effect and results confusion matrices by using combined features are shown in Fig. 6.

Observed from the matrices, the motion jogging is easily confused with running and walking. That is because the similarity between these three actions leads to error classification, which also accords with our visual observation. Fig. 7 shows the confusion matrices by using only 3D SIFT feature where the broken line graphs are with both features in SC2 and SC3 (the broken line doesn′t represent rate trend, but it can clearly contrast recognition rates). Compared with the corresponding confusion matrices of combined features (3D SIFT+PDI) in Fig. 6, it is noted that the identifiability of 3D SIFT to the confusing actions is improved by combining PDI, and the average recognition rate is respectively raised by 3% (SC2) and 6% (SC3). It is also verified that the proposed combined features are stable and adaptable.

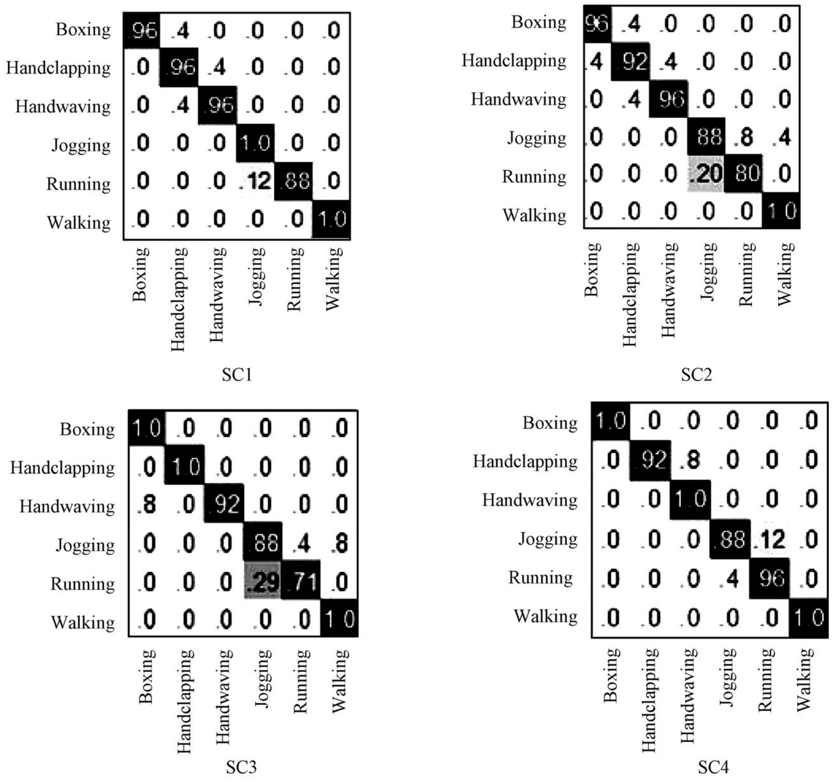
4.4 Testing results in mixed scenarios

From the analysis of experiments in portion scenarios in Section 4.3, 3D SIFT in combination with PDI has more advantages. Therefore, in this section, we test the combined features in mixed scenarios to validate the feasibility of our approach. The testing process still uses the leave-one-out cross validation method.

As shown in Fig. 8, since SC1 and SC4 are simple and stable relative to other mixed scenarios, they obtained the best recognition rate of 97%. The easily confused actions in the above mixed scenarios are still between jogging and running. Summarizing all the mixed two scenarios, the average recognition rate reaches 94.10%.

In order to make diversity of scenarios, we extend the number of mixed scenarios to three, e.g., SC123 (SC1+SC2+SC3) represents the mixture of scenario 1 ,2 and 3. Finally, we used all scenarios SC1234 (SC1 + SC2 + SC3 + SC4) to test our approach. The results are shown in Fig. 9. The confusion matrix of action recognition from the mixed multiple scenarios remained at about 94%. In order to more clearly verify the adaptability of our approach to mixed scenarios and make a comparison, we drew the results in the form of broken line graph (Confusion SC in Fig. 9). The proposed approach in this paper can get the better recognition rate even in a complex confusion of scenarios. Furthermore, it is also found that the actions can be captured in a stable scenario as action training samples, then those samples can be utilized for action recognition in unstable environment.

The comparison of performances between the proposed method and the recent related works based on KTH dataset are shown in Table 3. These works are all related to the local spatio-temporal feature. It is worth noting that our method outperforms all the other state of the art methods.



	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	96	.4	.0	.0	.0	.0
Handclapping	.0	96	.4	.0	.0	.0
Handwaving	.0	.4	96	.0	.0	.0
Jogging	.0	.0	.0	1.0	.0	.0
Running	.0	.0	.0	.12	88	.0
Walking	.0	.0	.0	.0	.0	1.0

SC1

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	96	.4	.0	.0	.0	.0
Handclapping	.4	92	.4	.0	.0	.0
Handwaving	.0	.4	96	.0	.0	.0
Jogging	.0	.0	.0	88	.8	.4
Running	.0	.0	.0	.20	80	.0
Walking	.0	.0	.0	.0	.0	1.0

SC2

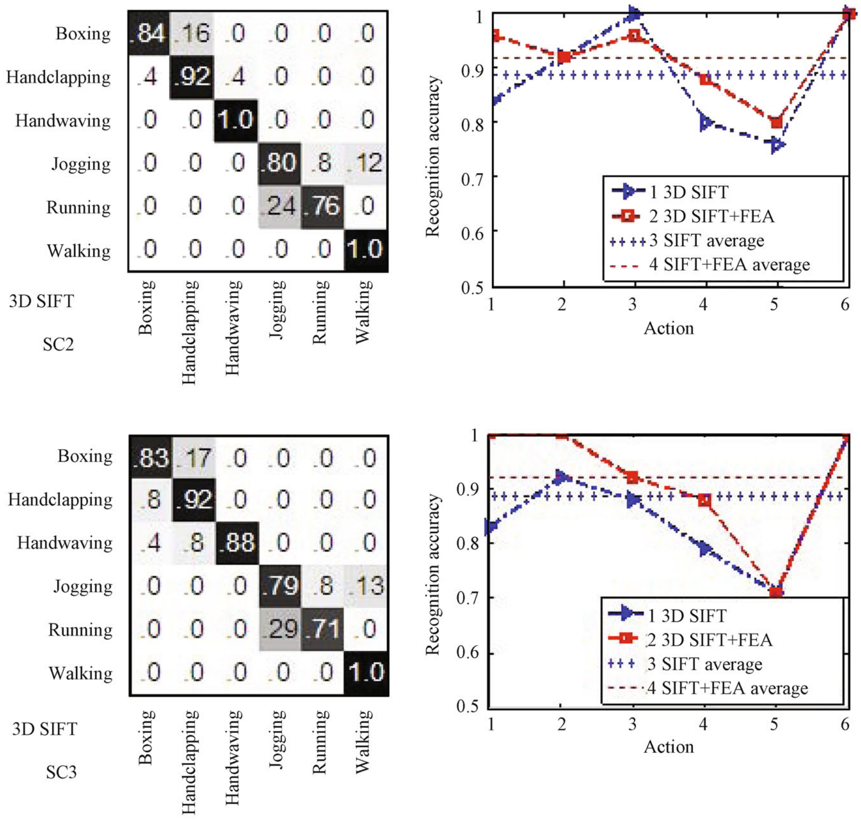
	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	1.0	.0	.0	.0	.0	.0
Handclapping	.0	1.0	.0	.0	.0	.0
Handwaving	.8	.0	92	.0	.0	.0
Jogging	.0	.0	.0	88	.4	.8
Running	.0	.0	.0	.29	71	.0
Walking	.0	.0	.0	.0	.0	1.0

SC3

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	1.0	.0	.0	.0	.0	.0
Handclapping	.0	92	.8	.0	.0	.0
Handwaving	.0	.0	1.0	.0	.0	.0
Jogging	.0	.0	.0	88	.12	.0
Running	.0	.0	.0	.4	96	.0
Walking	.0	.0	.0	.0	.0	1.0

SC4

Fig. 6 Confusion matrix of 3D SIFT+PDI recognition in portion scenario



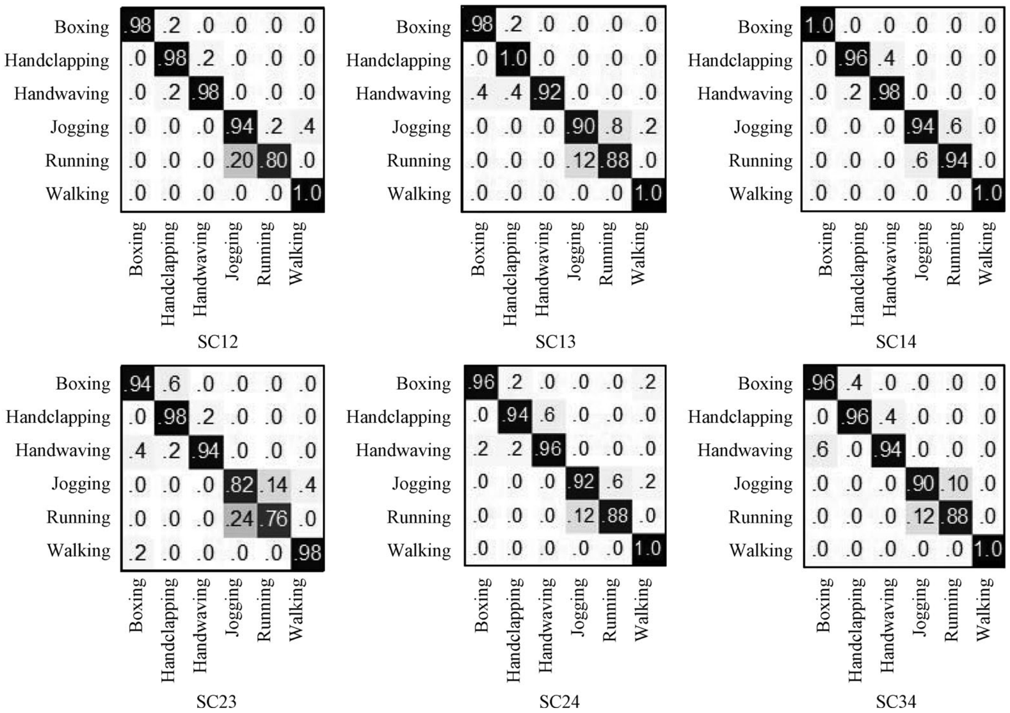
3D SIFT SC2

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.84	.16	.0	.0	.0	.0
Handclapping	.4	.92	.4	.0	.0	.0
Handwaving	.0	.0	1.0	.0	.0	.0
Jogging	.0	.0	.0	.80	.8	.12
Running	.0	.0	.0	.24	.76	.0
Walking	.0	.0	.0	.0	.0	1.0

3D SIFT SC3

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.83	.17	.0	.0	.0	.0
Handclapping	.8	.92	.0	.0	.0	.0
Handwaving	.4	.8	.88	.0	.0	.0
Jogging	.0	.0	.0	.79	.8	.13
Running	.0	.0	.0	.29	.71	.0
Walking	.0	.0	.0	.0	.0	1.0

Fig. 7 Contrast graph and the confusion matrices with 3D SIFT



SC12

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.98	.2	.0	.0	.0	.0
Handclapping	.0	.98	.2	.0	.0	.0
Handwaving	.0	.2	.98	.0	.0	.0
Jogging	.0	.0	.0	.94	.2	.4
Running	.0	.0	.0	.20	.80	.0
Walking	.0	.0	.0	.0	.0	1.0

SC13

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.98	.2	.0	.0	.0	.0
Handclapping	.0	1.0	.0	.0	.0	.0
Handwaving	.4	.4	.92	.0	.0	.0
Jogging	.0	.0	.0	.90	.8	.2
Running	.0	.0	.0	.12	.88	.0
Walking	.0	.0	.0	.0	.0	1.0

SC14

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	1.0	.0	.0	.0	.0	.0
Handclapping	.0	.96	.4	.0	.0	.0
Handwaving	.0	.2	.98	.0	.0	.0
Jogging	.0	.0	.0	.94	.6	.0
Running	.0	.0	.0	.6	.94	.0
Walking	.0	.0	.0	.0	.0	1.0

SC23

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.94	.6	.0	.0	.0	.0
Handclapping	.0	.98	.2	.0	.0	.0
Handwaving	.4	.2	.94	.0	.0	.0
Jogging	.0	.0	.0	.82	.14	.4
Running	.0	.0	.0	.24	.76	.0
Walking	.2	.0	.0	.0	.0	.98

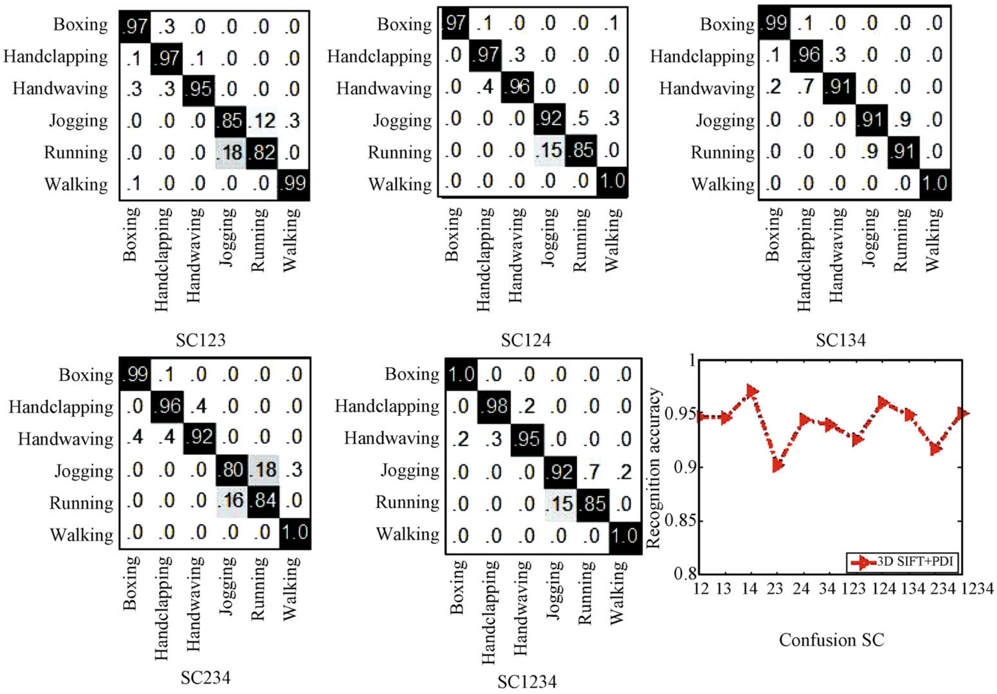
SC24

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.96	.2	.0	.0	.0	.2
Handclapping	.0	.94	.6	.0	.0	.0
Handwaving	.2	.2	.96	.0	.0	.0
Jogging	.0	.0	.0	.92	.6	.2
Running	.0	.0	.0	.12	.88	.0
Walking	.0	.0	.0	.0	.0	1.0

SC34

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.96	.4	.0	.0	.0	.0
Handclapping	.0	.96	.4	.0	.0	.0
Handwaving	.6	.0	.94	.0	.0	.0
Jogging	.0	.0	.0	.90	.10	.0
Running	.0	.0	.0	.12	.88	.0
Walking	.0	.0	.0	.0	.0	1.0

Fig. 8 Confusion matrix of two scenarios



	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.97	.3	.0	.0	.0	.0
Handclapping	.1	.97	.1	.0	.0	.0
Handwaving	.3	.3	.95	.0	.0	.0
Jogging	.0	.0	.0	.85	.12	.3
Running	.0	.0	.0	.18	.82	.0
Walking	.1	.0	.0	.0	.0	.99

SC123

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.97	.1	.0	.0	.0	.1
Handclapping	.0	.97	.3	.0	.0	.0
Handwaving	.0	.4	.96	.0	.0	.0
Jogging	.0	.0	.0	.92	.5	.3
Running	.0	.0	.0	.15	.85	.0
Walking	.0	.0	.0	.0	.0	1.0

SC124

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.99	.1	.0	.0	.0	.0
Handclapping	.1	.96	.3	.0	.0	.0
Handwaving	.2	.7	.91	.0	.0	.0
Jogging	.0	.0	.0	.91	.9	.0
Running	.0	.0	.0	.9	.91	.0
Walking	.0	.0	.0	.0	.0	1.0

SC134

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	.99	.1	.0	.0	.0	.0
Handclapping	.0	.96	.4	.0	.0	.0
Handwaving	.4	.4	.92	.0	.0	.0
Jogging	.0	.0	.0	.80	.18	.3
Running	.0	.0	.0	.16	.84	.0
Walking	.0	.0	.0	.0	.0	1.0

SC234

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	1.0	.0	.0	.0	.0	.0
Handclapping	.0	.98	.2	.0	.0	.0
Handwaving	.2	.3	.95	.0	.0	.0
Jogging	.0	.0	.0	.92	.7	.2
Running	.0	.0	.0	.15	.85	.0
Walking	.0	.0	.0	.0	.0	1.0

SC1234

Confusion SC

Fig. 9 Confusion matrices in mixed scenarios

Table 3 Comparison with related works in recent years

Literature	Method	Accuracy
[14]	3D SIFT BOW+pLSA	83.33%
[19]	3D Gradients+SVM	91.4%
[22]	Interest point clouds+NNC	93.17%
[24]	HOG3D+ SVM	92.7%
Our approach	3D SIFT+PDI + SVM	94.92%

5 Conclusions

This paper proposed a novel video descriptor by combining local spatio-temporal feature and global positional distribution information of interest points. Considering that the distribution of interest points contains rich motion information and also reflects the key position in human action, we combine the 3D SIFT with PDI to achieve a more complete representation of the human action. In order to obtain compact description and efficient computation, the combined features are processed by dimension reduction, feature quantization and feature selection. Eventually, compared with the previous works of 3D SIFT descriptor, the proposed approach further improved the recognition rate. In the future work, our approach will be applied to more complex datasets and applications and we will provide an additional performance improvement[25].

References

[1] H. J. Seo, P. Milanfar. Action recognition from one example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 867–882, 2011.

[2] D. Weinland, R. Ronfard, E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[3] X. Ji, 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, pp. 13–24, 2010.

[4] X. Li. HMM based action recognition using oriented histograms of optical flow field. *Electronics Letters*, vol. 43, no. 10, pp. 560–561, 2007.

[5] S. Ali, M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288–303, 2007.

[6] M. Hahn, L. Krüger, C. Wöhler. 3D action recognition and long-term prediction of human motion. In *Proceedings of the 6th International Conference on Computer Vision Systems, Lecture Notes in Computer Science*, Springer, Santorini, Greece, vol. 5008, pp. 23–32, 2008.

[7] F. Jiang, Y. Wu, A. K. Katsaggelos. A dynamic hierarchical clustering method for trajectory-based unusual video event detection. *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 907–913, 2009.

[8] H. Y. Zhou, H. S. Hu, H. H. Liu, J. S. Tang. Classification of upper limb motion trajectories using shape features. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 6, pp. 970–982, 2012.

[9] X. B. Cao, B. Ning, P. Yan, X. L. Li. Selecting key poses on manifold for pairwise action recognition. *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 168–177, 2012.

[10] A. A. Chaaraoui, P. Climent-Pérez, F. Flórez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[11] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[12] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie. 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*, IEEE, Beijing, China, pp. 65–72, 2005.

[13] I. Laptev, T. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proceedings of the lst Workshop on Spatial Coherence for Visual Motion Analysis*, Springer, vol. 3667, pp. 91–103, 2006.

[14] J. C. Niebles, H. C. Wang, F. F. Li. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[15] J. W. Zhu, J. Qi, X. B. Kong. An improved method of action recognition based on sparse spatio-temporal features. In *Proceedings of Artificial Intelligence: Methodology, Systems, and Applications*, Springer, vol. 7557, pp. 240–245, 2012.

[16] P. Liu, J. Wang, M. She, H. H. Liu. Human action recognition based on 3d SIFT and LDA model. In *Proceedings of IEEE Workshop on Robotic Intelligence in Informationally Structured Space*, IEEE, Paris, France, pp. 12–17, 2011.

[17] X. H. Jiang, T. F. Sun, B. Feng, C. M. Jiang. A space-time surf descriptor and its application to action recognition with video words. In *Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, Shanghai, China, vol. 3, pp. 1911–1915, 2011.

[18] P. Scovanner, S. Ali, M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia*, ACM, New York, USA, pp. 357–360, 2007.

[19] A. Kläser, M. Marszałek, C. Schmid, L. LEAR. A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of British Machine Vision Conference*, BMVA Press, UK, pp. 1–10, 2008.

[20] G. Willems, T. Tuytelaars, L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of European Conference on Computer Vision*, Springer, France, vol. 5303, pp. 650–663, 2008.

[21] F. Li, J. X. Du. Local spatio-temporal interest point detection for human action recognition. In *Proceedings of the 5th International Conference on Advanced Computational Intelligence*, IEEE, Nanjing, China, pp. 579–582, 2012.

[22] M. Bregonzio, S. G. Gong, T. Xiang. Recognising action as clouds of space-time interest points. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, IEEE, Miami, USA, pp. 1948–1955, 2009.

[23] C. C. Chang, C. J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, Article 27, 2011.

[24] S. Umakanthan, S. Denman, S. Sridharan, C. Fookes, T. Wark. Spatio temporal feature evaluation for action recognition. In *Proceedings of International Conference on Digital Image Computing: Techniques and Applications*, Fremantle, Western Australia, pp. 1–8, 2012.

[25] J. M. Chaquet, E. J. Carmona, A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

Xiao-Fei Ji received the M. Sc. degree from the Liaoning Shihua University, China in 2003, and Ph. D. degree from University of Portsmouth, UK in 2010. From 2003 to 2012, she was a lecturer at School of Automation of Shenyang Aerospace University, China. From 2013, she became an associate professor at Shenyang Aerospace University. She is the leader of National Natural Science Fund Project (No. 61103123) and main group member of 6 national and local government projects. She is IEEE member, she 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.

E-mail: jixiaofei7804@126.com (Corresponding author)

Qian-Qian Wu received the B. Eng. degree from Langfang Teacher's College China in 2011, and received the M. Sc. degree from the School of Automation, Shenyang Aerospace University, China in 2013. She is currently an engineer in an aeronautical enterprise.

Her research interests include video analysis, human action modeling and recognition.

E-mail: 847559648@qq.com

Zhao-Jie Ju received the B. Sc. degree in automatic control and the M. Sc. degree in intelligent robotics from Huazhong University of Science and Technology, China in 2005 and 2007, respectively. He received the Ph. D. degree in intelligent robotics at University of Portsmouth, UK in 2010. He is currently a lecturer at School of Computing, University of Portsmouth, UK. He previously held research appointments in the Department of Computer Science, University College London, UK, and Intelligent Systems and Biomedical Robotics Group, University of Portsmouth, UK.

His research interests include machine intelligence, robot learning, pattern recognition, and their applications to robotic/prosthetic hand control and human-robot interaction.

E-mail: zhaojie.ju@port.ac.uk

Yang-Yang Wang received the M. Sc. degree from Shenyang Aerospace University, China in 2006. She is currently a Ph. D. candidate at College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, China. She has published over ten research papers in this research direction.

Her research interests include human action modeling and recognition.

E-mail: wyy2004101@163.com