International Journal of Automation and Computing X(X), X X, X-X
DOI: XXX

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

Xiaofei Ji[1] Qianqian Wu[1] Zhaojie Ju[2] Yangyang Wang[1]

[1]School of Automation, Shenyang Aerospace University, Shenyang 110136, PRC

[2]School of Creative Technologies ,University of Portsmouth, Portsmouth PO5 4BP,UK

Abstract: Most of the existed action recognition methods mainly utilize spatio-temporal descriptors of single interest point 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 points detection method. Then 3-dimensional scale-invariant feature transform (3D SIFT) descriptors are extracted for every interest point. In order to obtain compact description and efficient computation, Principal Component Analysis (PCA) method is utilized twice on the 3D SIFT descriptors of single-frame and multi-frame. 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 Support Vector Machine (SVM) recognition algorithm on the public KTH dataset. The testing results showed that the recognition rate has been significantly improved. Meantime, the test results verified the proposed features can more accurately describe human motion with high adaptability to scenarios.

Keywords: action recognition, spatio-temporal interest points, 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 sports events, and more [1]. Due to the complexity of the action, such as different body wearing and habits leading to different observation of the same action, the camera movement in the external environment, illumination change, shadows, viewpoint and so on, these influence of factors make action recognition still a challenging project[2, 3].

The representation of human motion in video sequences is crucial in 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 and so on. At present, most commonly used features in action recognition are based on motion, such as optical flow[4, 5], motion trajectory[6, 7, 8] *etc.*, or based on the appearance shape, for example silhouette contour[9, 10] *etc.*. The former features are greatly influenced by illumination, 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, 13, 14, 15].

Spatio-temporal interest points are those points where the local neighborhood has a significant variation in both the spatial and the temporal domain. Most of local spatio-temporal feature descriptors are the extension of the information extracted form previously 2D space to 3D spatio-temporal based on the interest points detected by Laptev[13] or Dollar[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 SIFT[16], SURF[17] and so on, which have advantages of scale, affine, view and rotation invariance. Dollar[12] applied the cuboids and selected smooth gradient as a descriptor. Later, Scovanner[18] *et al.* put forward the 3D SIFT to calculate spatio-temporal gradients direction histograms for each pixel within its neighborhood. Another extension to the SIFT was proposed by Klaser[19] *et al.*, based on a histogram of 3D gradient orientations, where gradients are computed using an integral video representation. Williems[20] *et al.* 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, Dollar[12] adopted the PCA algorithm to reduce feature dimension and finally made use of nearest neighbor classifier and SVM to recognize human actions on KTH dataset. Niebles[14] considered 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[21] got interest points from Harris detector and then extracted 3D SIFT descriptor, in recognition process he made use of SVM with leave-one-out method and also did the experiment on 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 u-

Manuscript received date; revised date
The Project supported by the National Natural Science Foundation of China (No. 61103123) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

tilized local spatio-temporal descriptors of single interest point 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 sports information also changes accordingly. In addition, it can't simply rely on the local feature of spatio-temporal interest points to represent the target motion when the motion lacks time dimension information. Bregonzio[22] *et al.* defined a set of features which reflect the interest points distribution based on different temporal scales. In their study, global spatio-temporal distribution of interest points is studied but the excellent performance of local descriptor is also abandoned. When the influence factors interfere body movement, it cannot reliably represent the action only depending on the global information. These above experiments are all conducted on the KTH dataset for the recognition test, but the adaptability of feature applied in different scenarios on the KTH is not studied and discussed. Moreover, despite Dollar[12] detection method's popularity, it tends to generate spurious detection background area surrounding 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. That is a combination of 3D SIFT descriptor and the spatio-temporal distribution information based on interest points. First step, an improved detecting method[22] is used to detect spatio-temporal interest points, different from Dollar's method, effectively avoiding 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 perfect combination, the dimension reduction issue is performed on descriptor twice, which is based on single-frame and multi-frame. Then the processed descriptor is combined with positional distribution information of interest points. Finally, the combined features are quantified and selected to obtain more concise feature descriptors. 3D SIFT descriptor contains human body posture information and motion dynamic information, 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 changes in human body shape and motion directions *etc.*. So the feature has good adaptability and robustness in complex motion scenarios. The positional distribution information of interest points reflects motion global information by using various location and ratio relationship of the two areas of human body movement and interest points distribution. Different from previous methods of describing the appearance and shape in space, this paper don't directly pick up the shape information. So when the human appearance and shape changes, the location and ratio relationship of the two areas are not directly affected. Therefore, the positional distribution information of interest points proposed by this paper is more adaptive for motion description in space. Finally the proposed motion descriptor by combining the above mentioned local feature with global information is tested by using to SVM recognition algorithm on the public dataset of KTH. Furthermore, the adaptability of proposed method is discussed by testing in each different and mixed scenarios of KTH dataset. By comparing with the related and similar research works in recent years, the results verified the proposed method is better with 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 in this paper. 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. And Section 4 gives experimental results and analysis. Finally, Section 5 concludes the paper.

2 Interest Points Detection

In computer vision, interest points represent the location which has severe changes in space and time dimensions and considered to be salient or descriptive for the action captured in a video. Among various interest points detection methods, the most widely used for action recognition is the one proposed by Dollar[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.

Dollar'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 the Dollar detector, in particular, the Gabor filtering which do the feature extraction only on the time axis, ignoring the dynamic movement in the prospects. To overcome these shortcomings, we utilize a different interest points 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 follow three steps:

1. frame differencing for focus of human action and region of interest detection, as shown in Fig.1(b).

2. utilizing 2D Gabor filter for generating five different directions $(0°, 22°, 45°, 67°, 90°)$ filter templates, the example presented in Fig.1(b).

3. Filtering on the detected regions of interest got in step 1 , using 2D Gabor filters with five different orientations obtained by step 2 in both the spatial and temporal domains to give a combined filter response.

Fig.1(c) shows examples of our interest point detection results (green circle points) and the Dollar's[12] (red square points). It is evident that the detected interest points are much more meaningful and descriptive compared to those detected using the Dollar detector.

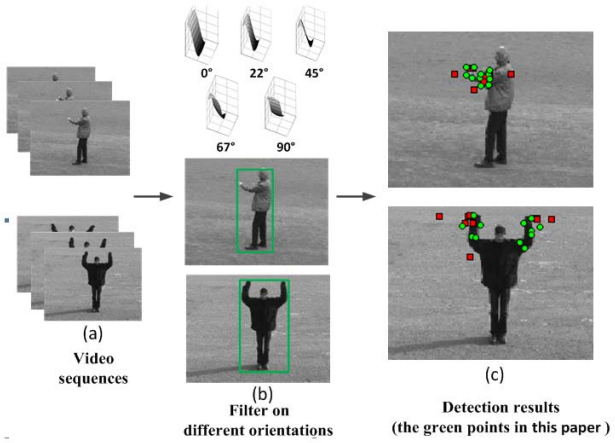

Figure 1 Interest points detection process

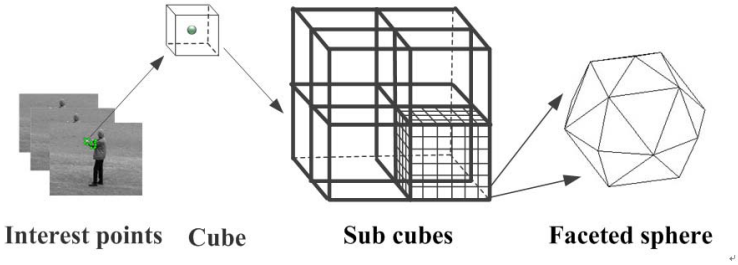

Figure 2 Description process of 3D SIFT

3 Action Representation

3.1 3D SIFT Descriptor

For calculating the SIFT descriptor, firstly 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 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 $12 \times 12 \times 12$ pixel size cube is divided into $2 \times 2 \times 2$ sub cubes, as shown in Fig.2.

3-dimensional gradient magnitude at (x, y, t) is computed by Eq.1.

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

The $L_x = L(x+1, y, t) - L(x-1, y, t)$, $L_y = L(x, y+1, t) - L(x, y-1, t)$, $L_t = L(x, y, t+1) - L(x, y, t-1)$ stand for the gradient from x, y, t direction 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 gradient histogram by using the Eq. 2.

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

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

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

3.2 Positional Distribution Information of Interest Points

In terms of interest points in each frame, the positional distribution information is closely related to body motion, it also reflects the amplitude range of action, relevance of human body location and motion parts 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, 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) box respectively. Then the related positional distribution information is calculated with these two areas and expressed by $PDI = [D_{ip}, R_{ip}, R_{ren}, Vertic_{dist}, Orizon_{dist}, W_{ratio}, H_{ratio}, Over_{lap}]$. A calculation method of the features are shown in Table 1.

In order to remove the influence of several stray individual points on the overall feature, stray filtering process is utilized on all the interest points before extracting the

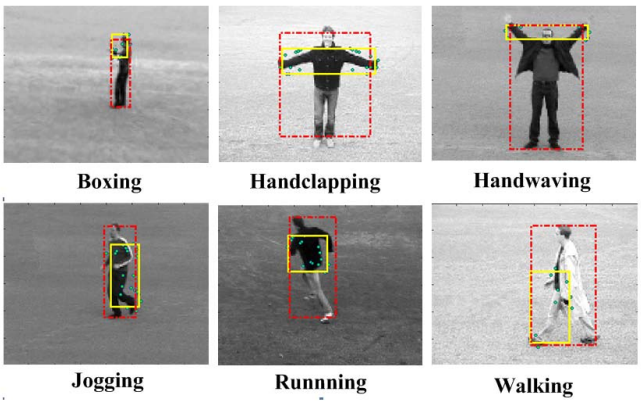

Figure 3 Distribution of interest points

Table 1 Calculation method of PDI feature

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

above positional feature. The points whose distance to the region centroid 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 attribute of the motion.

3.3 Motion Features

In section 3.1, the 3D SIFT descriptor of each interest point is 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 spatio-temporal information in this frame. The dimension of feature is so high that it cannot reasonable combine with 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 datas and make features more concise, above combined feature are processed by using quantization and selection. The following five steps are listed as:

1. *Single frame dimension reduction:* Principle component analysis(PCA) 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 component of all 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 the N is chosen as 20 in our experiment.

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 frames to set $M \times 256$ (M for total number of this video frames) to $M \times 50$.

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 gets 58 dimension features for each frame (3D SIFT+ PDI).

4. *Feature quantization:* The linear quantization is utilized on $M \times 58$ dimension feature of each video to a histogram containing $n(n < M)$ bins, thus it turned to be $n \times 58$ dimension feature.

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.

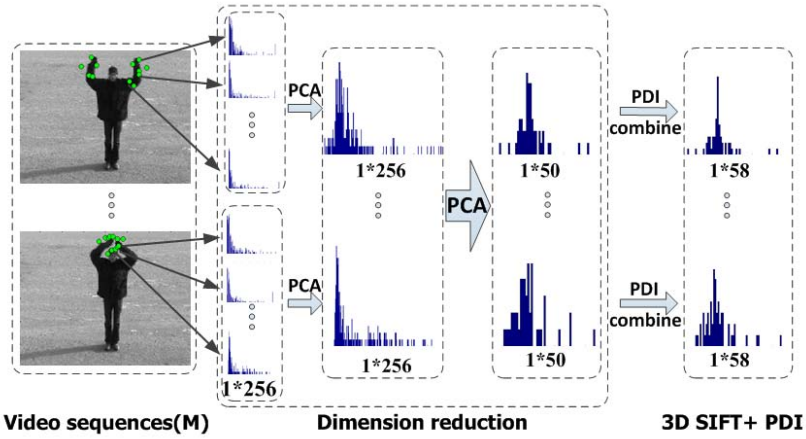

Figure 4 The description of motion feature

4 Algorithm Verification and Results Analysis

In this section, experiments are performed on the K-TH 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 paper.

4.1 Recognition Algorithm

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

According to the theory, SVM is developed from the theory of Structural Risk Minimization, as shown in Eq.3.

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

For a given sample set x_i, $y_i \in \{-1,1\}$,$i = 1,\cdots n$, where $x_i \in R^N$ 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 linear 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 KKT conditions, and the dual function is written as:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(\phi(x_i),\phi(x_j))$$
$$C \geq \alpha_i \geq 0, i = 1,\cdots,n \qquad \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{4}$$

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

$$y(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i y_i k(x_i,x) - b) \tag{5}$$

Instead of establishing SVMs between one against the rest types, we adopt the method of one against one to establish a SVM between any two categories. The current sample belongs to which category determined by the decision function, and its final type is decided to the category with 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 respectively performed by 25 different persons: boxing, hand clapping, hand waving, jogging, running, 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 is shown in Fig.5. Apparently, due to the change of camera zooming situation, the size of human body change a lot in SC2(captured in t1 and t2). Furthermore person in SC3 put on different coat, or wear a hat or a bag leading to larger changes in body appearance.

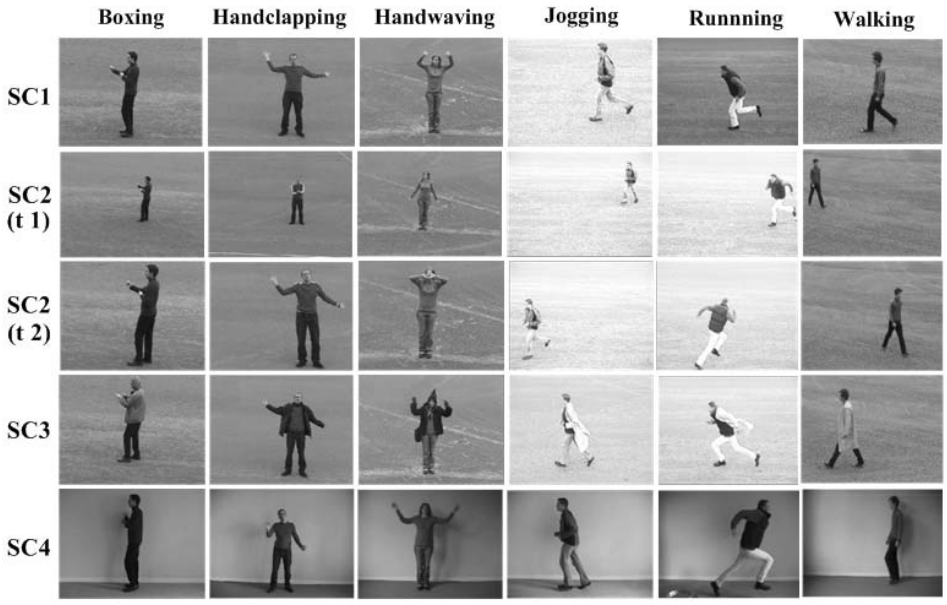

Figure 5 The description of motion feature

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. Through the tests in various scenarios with the features proposed in this paper to check out the reliability for motion description and the adaptability to scenarios. Leave-one-out cross validation method is adopted throughout the process, in turns using six action of each actor as test samples, and the rest of all the actions as the training, circulation continued until all actions are completed testing. The experimental results are shown in Table 2.

Table 2 Testing results in portion scenario

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

The experimental results show that the feature of 3D sift has the 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). Although emerging the influences of motion direction changes and indoor lighting, the 3D SIFT (after the process of dimension reduction and quantization) still is a good feature description for motion. It also shows that 3D sift has good adaptability and robustness to motion direction, position, speed and so on.

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, walking. In SC3, human body shape changes with different wearing, and the phenomenon of inhomogenous 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 rate greatly increased by using combined features(3D SIFT+PDI) compared to 3D SIFT. The recognition effect and results confusion matrixes by using combined features are shown in Fig.6.

Observed from the matrixes, the motion jogging is easily confused with running and walking. That is because the similarity between these three actions leads to the error classification, this also accords with our visual observation. Fig.7 shows the confusion matrixes by using only 3D SIFT feature and broken line graph with both features in SC2 and SC3 (broken line don't represent rate trend, while it can clearly contrast recognition rate high or low). Compared with the corresponding confusion matrixes of combined features(3D SIFT+PDI)in Fig.6, it is note that the identifiability of 3D SIFT to the confusing actions is improved by combining PDI, and the average recognition rate respectively raised 3%($SC2$)and 6%($SC3$). It also verified that the proposed combined features is 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 combined features in mixed scenarios to validate the feasibility of our approach. The testing process still uses leave-one-out cross validation method.

As shown in Fig.8, first of all, mix two scenarios, such as stable outdoor SC1 respectively mixed with scale variations SC2, different clothes wearing SC3 mixed with lighting SC4.

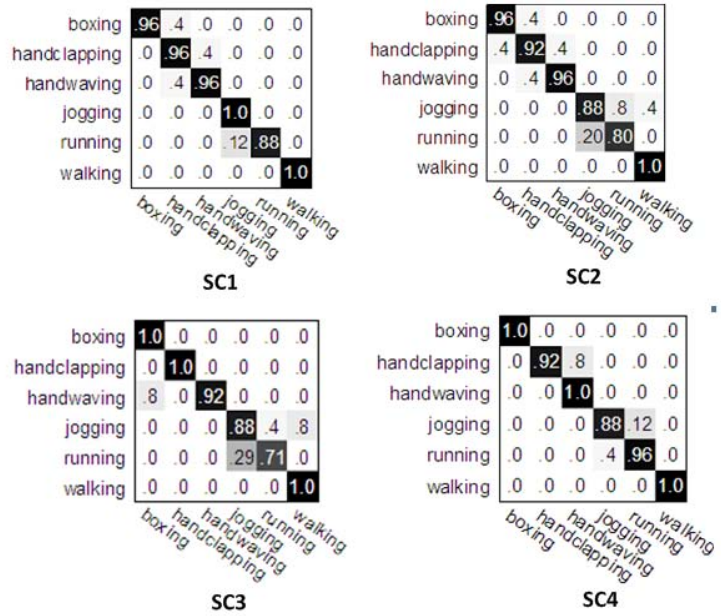

Figure 6 Confusion matrix of 3D SIFT+PDI recognition in portion scenario

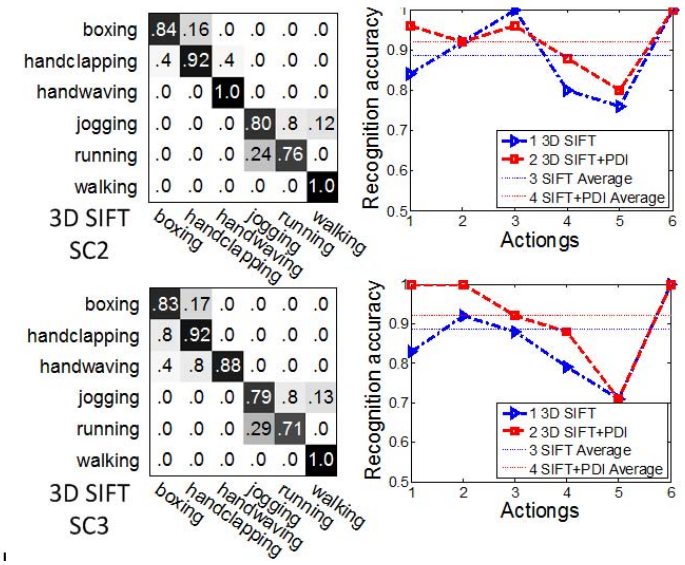

Figure 7 Contrast graph and the confusion matrixes with 3D SIFT

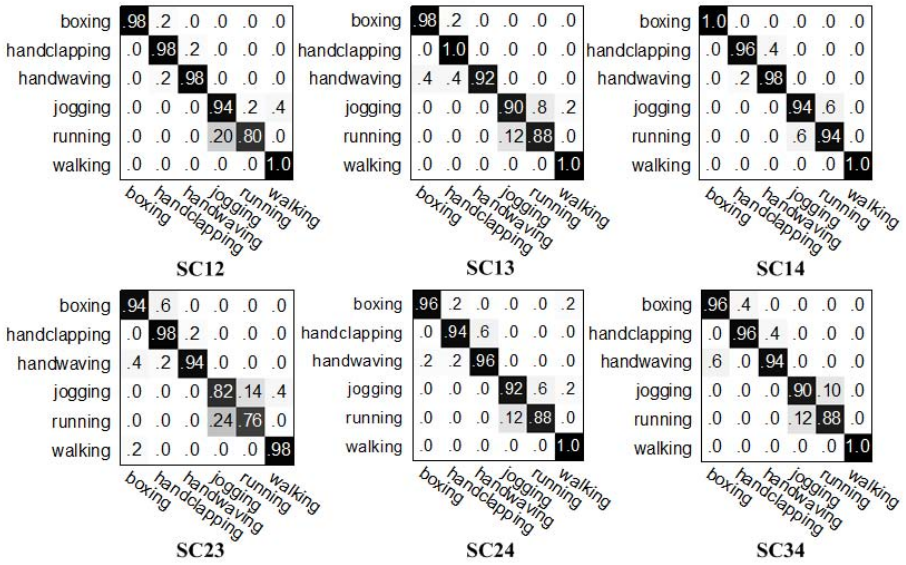

Figure 8 Confusion matrix of two scenarios

Through different complexity environment of two mixed scenarios to test our approach. As shown in Fig8. Due to SC1 and SC4 are relatively simple and stable than other mixed scenarios, so it obtained the best recognition rate 97%. The easily confused actions in 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, for example SC123(SC1+ SC2+ SC3), SC134(SC1+ SC3+ SC4)mixed three scenarios, finally used all scenarios SC1234 (SC1 and SC2 + SC3 + SC3) 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 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 comparisons of performance 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 of other state of art methods.

Table 3 Comparison with related work in recent years

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

5 Conclusion

This paper proposed a novel video descriptor by combining local spatio-temporal feature and global positional distribution information (PDI) of interest points. Considering that the distribution of interest points contains rich motion information and also reflects the key position in human action, so 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 previous works of 3D SIFT descriptor, the proposed approach further improved the recognition rate. In future work, our approach will be applied to more complex datasets and applications and 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, IEEE Transactions on 33 (5) (2011) 867–882.

[2] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Computer Vision and Image Understanding 115 (2) (2011) 224–241.

[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, 40 (1) (2010) 13–24.

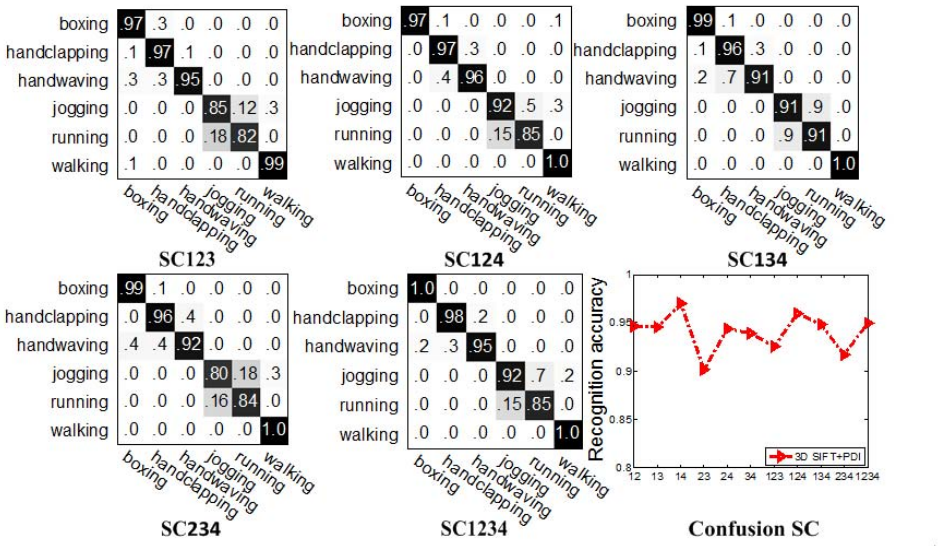

Figure 9 Confusion matrixes in mixed scenarios

[4] X. Li, Hmm based action recognition using oriented histograms of optical flow field, Electronics Letters 43 (10) (2007) 560–561.

[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, 32 (2) (2010) 288–303.

[6] M. Hahn, L. Krüger, C. Wöhler, 3d action recognition and long-term prediction of human motion, Computer Vision Systems (2008) 23–32.

[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, 18 (4) (2009) 907–913.

[8] H. Zhou, H. Hu, H. Liu and J. Tang, Classification of upper limb motion trajectories using shape features, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews (2012),42 (6) 970-982

[9] X. Cao, B. Ning, P. Yan, X. Li, Selecting key poses on manifold for pairwise action recognition, IEEE Transactions on Industrial Informatics, 8 (1) (2012) 168–177.

[10] A. A. Chaaraoui, P. Climent-Pérez, F. Flórez-Revuelta, Silhouette-based human action recognition using sequences of key poses, Pattern Recognition Letters.

[11] R. Poppe, A survey on vision-based human action recognition, Image and vision computing 28 (6) (2010) 976–990.

[12] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.

[13] I. Laptev, T. Lindeberg, Local descriptors for spatio-temporal recognition, in: Spatial Coherence for Visual Motion Analysis, Springer, 2006, pp. 91–103.

[14] J. C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision 79 (3) (2008) 299–318.

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

[16] P. Liu, J. Wang, M. She, H. Liu, Human action recognition based on 3d sift and lda model, in: 2011 IEEE Workshop on Robotic Intelligence In Informationally Structured Space (RiiSS), 2011, pp. 12–17.

[17] X. Jiang, T. Sun, B. Feng, C. Jiang, A space-time surf descriptor and its application to action recognition with video words, in: 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. 3, 2011, pp. 1911–1915.

[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, 2007, pp. 357–360.

[19] A. Kläser, M. Marszałek, C. Schmid, L. LEAR, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference, 2008, pp. 1–10.

[20] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: European Conference on Computer Vision, Springer, 2008, pp. 650–663.

[21] F. Li, C. Xiamen, J. Du, Local spatio-temporal interest point detection for human action recognition, in: IEEE 5th International Conference on Advanced Computational Intelligence, 2012, pp. 1–10.

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

[23] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 1–39.

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

[25] J.M. Chaquet, E.J. Carmona*et al*, A survey of video datasets for human action and activity recognition, Computer Vision and Image Understanding, 117 (6) (2013)633–659

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 2003 to 2012, she was the Lecturer at School of Automation of Shenyang Aerospace University. 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. She is the leader of National Natural Science Fund Project(Number: 61103123) and main group member of 6 National and Local Government Projects.

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

QianQian Wu received her B.Eng. degree from Langfang Teacher's College in 2011 and received her M.S. degrees from the school of automation, Shenyang Aerospace University, in 2013. She currently is an engineer in an aeronautical enterprise. Her research is focus on the video analysis,human action modeling and recognition. She has published 3 research papers in this research direction.

Zhaojie Ju received the B.S. in automatic control and the M.S. in intelligent robotics both from Huazhong University of Science and Technology, China, in 2005 and 2007 respectively, and the Ph.D. degree in intelligent robotics at the University of Portsmouth, UK, in 2010.

Dr Ju is currently a Lecturer in the School of Computing, University of Portsmouth, UK. He previously held research appointments in the Department of Computer Science, University College London and Intelligent Systems and Biomedical Robotics group, University of Portsmouth, UK. His research interests are in machine intelligence, robot learning, pattern recognition and their applications in robotic/prosthetic hand control.

Yangyang Wang received her M.S. degrees from the Shenyang Aerospace University, in 2006. She is currently a graduate student studying for Doctor degree in the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics. Her research is focus on the human action modeling and recognition. She has published over ten research papers in this research direction.