International Journal of Hybrid Information Technology
Vol.8, No. 5 (2015), pp. 165-176
http://dx.doi.org/10.14257/ijhit.2015.8.5.19

A Novel Action Recognition Method Based on Improved Spatio-Temporal Features and AdaBoost-SVM Classifiers

Xiaofei Ji, Lu Zhou and Qianqian Wu

School of Automation, Shenyang Aerospace University
Shenyang, China
Email: jixiaofei7804@126.com
School of Automation, Shenyang Aerospace University
Shenyang, China
School of Automation, Shenyang Aerospace University
Shenyang, China

Abstract

Most of existed action recognition methods based on spatio-temporal descriptors have ignored their spatial distribution information. However the spatial distribution information usually is very useful to improve the discriminative ability of the motion representation. An improved spatio-temporal is proposed in this paper by combining local spatio-temporal feature and global positional distribution information (FEA) of interest points. Furthermore, in order to improve the classifier's performance, an Adaboost-SVM method is utilized to recognize the human actions by using the proposed motion descriptor. The proposed recognition method is tested on the public dataset of KTH. The test results verified the proposed representation and recognition method can more accurately describe and recognize the human motion.

Keywords: *Action recognition, Spatio-temporal interest points, 3D SIFT, Positional distribution information, AdaBoost-SVM*

1. Introduction

Visual-based human action recognition has received considerable attention in computer vision during the past few years. The growing interest is due to a large number of real world applications, such as video surveillance, sport events analysis, human-computer interaction, and so on. However it remains challenging to recognize human actions performed by actors of different size, appearance and action habits [1, 2].

The representation of human motion in video sequences is crucial in action recognition. Ideally, the features should be robustness to small variations in appearance, background, and viewpoint and action execution. The representation of human motion can be divided into two categories: global representations and local representations [3]. The former encodes the region of interest (ROI) of a person as a whole. The common global representations are derived from optical flow [4, 5], silhouettes or edges [6, 7]. They are sensitive to noise, partial occlusion and variations in viewpoint. Local representation describes the observation as a collection of independent patches. Compared with the global representation, local features are somewhat invariant to changes in viewpoint, person appearance and partial occlusions. Due to their advantage, local spatio-temporal features based on interest points are more and more popular in action recognition [8, 9].

Spatio-temporal interest points are those points where the local neighborhood has a significant variation in both the spatial and the temporal domain. It is assumed that these locations are most informative for the recognition of human action. Up to now, many efforts have been devoted to the description of the spatio-temporal interest points. The

ISSN: 1738-9968 IJHIT

most common descriptions are SIFT [10], SURF [11] and so on, which have advantages of scale, affine, view and rotation invariance. In the process of recognition, Niebles [8] 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 [12] got interest points from Harris detector and then extracted 3D SIFTS descriptor, and then tested the features on KTH dataset by using SVM algorithm with leave-one-out method. The above recognition methods have achieved good recognition results, but most studies only stayed on the description of the interest points, and mainly utilized local spatio-temporal descriptors of single interest point ignoring its overall distribution information in the global space and time. Bregonzio[13] *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.

Based on the above discussions, a novel recognition method is proposed in this paper. On the one hand, a novel human motion representation is proposed by combining local and global information. That is a combination of 3D SIFT descriptor and the spatio-temporal distribution information based on interest points. 3D SIFT descriptor contains human body posture information and motion dynamic information [14]. It describes the local feature of action both in spatio and temporal dimension. 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. On the other hand, an Adaboost-SVM method is utilized to recognize the human actions by using the proposed motion descriptor. AdaBoost equally partitioned the whole classification into several layers. It constructed one non-linear SVM in each layer. Through satisfied combination and iterative weights, the algorithm focus on hard-classifier's to improve the classifier's performance. The whole framework of the proposed method is shown as Figure 1.

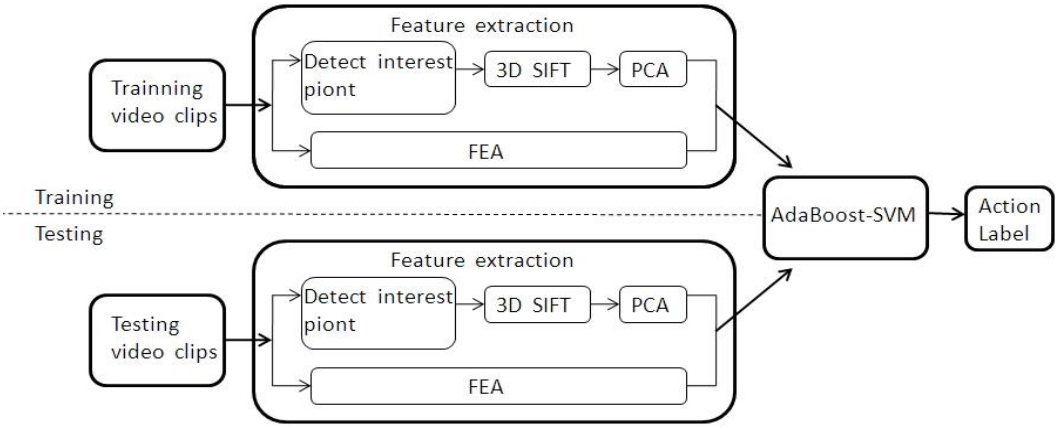

Figure 1. The Framework of the Proposed Recognition Method

Finally the proposed recognition method is tested on the public dataset of KTH. By comparing with the related and similar research works in recent years, the results verified the proposed method is better with recognition accuracy and adaptability.

2. Feature Extraction and Representation

The representation of human motion in video sequences is crucial in action recognition. A novel motion representation by combining 3D SIFTS descriptor of interest points and

the spatio-temporal distribution information based on interest points is proposed in this section.

2.1. Interest Point Feature Extraction

The interest point is referred to the changing point in spatio-temporal domain which can be distinguished from other points within a neighborhood. So the interest points can be used as a feature to represent the action sequence. Among various interest points detection methods, the most widely used for action recognition is the one proposed by Dollar [15]. Dollar's method for interest point's detection of human motion in video has certain effect, however, it is prone to false detection due to video shadow and noise, and spurious interest points are easy to occur in the background. So a different interest point detector is utilized [14]. The region of interest detected by the frame differencing algorithm is filtered using 2D Gabor filters from different orientation (0◦ , 22◦ , 45◦ , 67◦ and 90◦ orientations selected). Then combination of different orientation filtering responses is used for the final detection, as shown in Figure1 (a)-(d).

Then the 3D Scale-invariant Feature Transform (3D SIFT) descriptor [16] is utilized to represent the interest points in these frames. In this paper, the 12∗12∗12 pixel size cube is divided into eight 2∗2∗2 sub cubes, and 32 faceted spheres at 32 gradient directions is utilized in each sub cubes for descriptor. So the feature dimension of each sub cube is 32. The initial whole features of each point are 256 dimensions, as shown in Figure 1 (e).

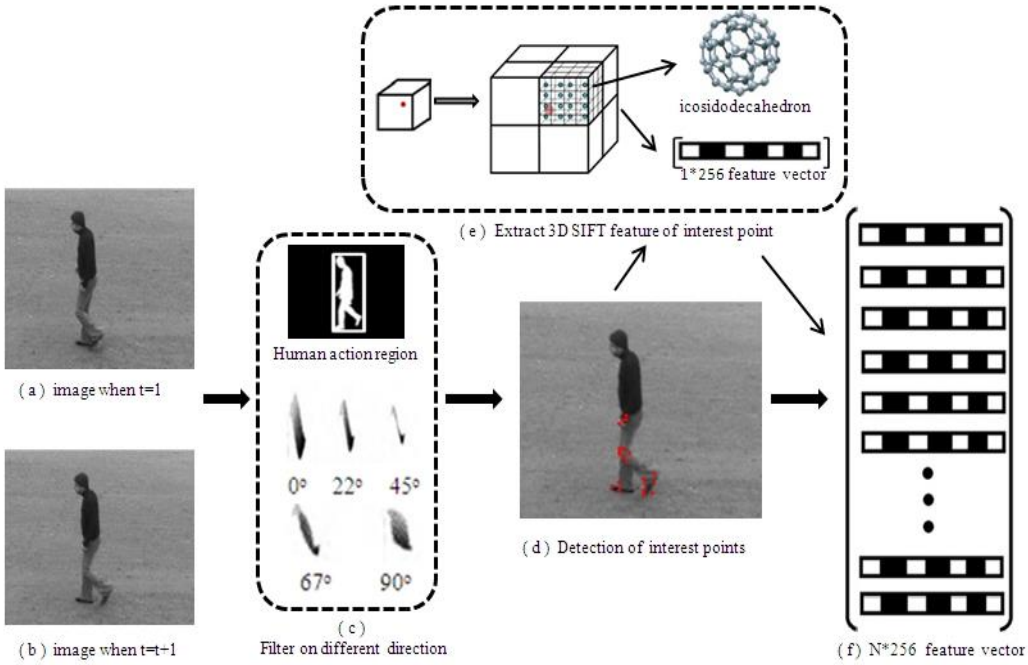

Figure 2. The Interest Point Detection and 3D SIFT Presentation

2.2. Interest Point Feature Dimension Reduction

The 3D SIFT descriptor of each interest point is 256 dimensions, if the number of interest points in each frame is N, the feature is N *256 dimensions to represent spatio-temporal information in each frame, as shown in Figure1 (f). Considering the dimension of feature is too high, so the dimension reduction is performed as following two steps:

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 *256 features can be reduced to 1 * 256 by gathering principal component of all descriptors for each frame.

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 *256 (M for total number of this video frames) to M *50.

2.3 Position Distribution Information of Interest Points

In terms of interest points in each frame, the positional distribution information is closely related to body motion. And it can reflect the amplitude range of action and the relation between the human body location and the motion parts region. So the position distribution information of interest points is extracted as another kind of action information. The specific process is described as below. As shown in Figure 2, the distribution region of interest points and body location region are detected in each frame. The distribution region and body location region are drawn with yellow (Y) and red (R) box respectively.

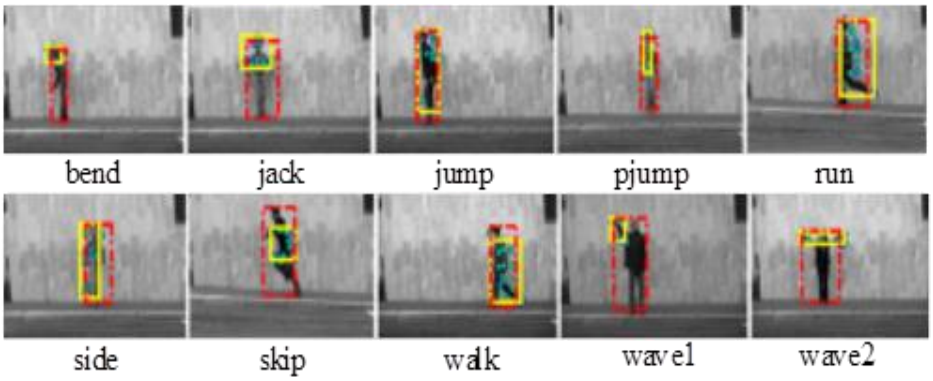

Figure 3. Distribution Information of Interest Points

Then the related positional distribution information is calculated with these two areas. The detail of the calculation method is defined as following:

FEA =[Dip; Rip; Rren; Verticdist; Orizondist; Wratio; Hratio; Overlap]

Dip: the total number of point normalized by the region Y

Rip: the height and width ratio of the region Y

Rren: the height and width ratio of the region R

Verticdist: the vertical distance between the geometrical center (centroid) of the region Y and the region R

Orizondist: the horizontal distance between the geometrical center (centroid) of the region Y and the region R

Wratio: the width ratio between the region Y and the region R

Overlap: the ratio by the amount of overlap and total width between the region Y and the region R

The positional distribution information of interest points is extracted from each frame to represent and reflect the whole attribute of the motion. The dimension of the FEA in every frame is 8.

2.4. The Combination of the Interest point feature and PDI

In order to improve the ability to distinguish of the feature, a novel feature is proposed by combing the spatio-temporal features(1*50 dimension) and the corresponding positional distribution information (1*8 dimension) of interest points in each frame,

finally gets 58 dimension features to represent the motion information in each frame (3D SIFT+ FEA).

3. Action Recognition Based on Adaboost-SVM Classifiers

AdaBoost has been widely used to improve the accuracy of any give learning algorithm. In this section, we focus on designing an algorithm to combine AdaBoost and SVM as weak classifiers to be used in human action recognition.

3.1. AdaBoost Algorithm [17-18]

AdaBoost algorithm is based on iterative algorithm. The core of the algorithm is to train multiple weak classifiers according to the sub training data. AdaBoost classifier is an ensemble strong classifier composed of many weak classifiers based on certain rules. AdaBoost constructs a composite classifier by sequentially training classifiers while putting more and more emphasis on certain patterns. AdaBoost classification can effectively exclude some unnecessary training sample data, and then locate the most critical training sample to improve the accuracy of training and save resource consumption.

The schematic diagram of AdaBoost algorithm is shown in Figure 4. AdaBoost algorithm achieves new training subset by calculating and adjusting the weight of each sample in the iteration process. Given a set of training sample *C*, AdaBoost maintains a probability distribution V_1. This distribution is initially uniform. AdaBoost algorithm calls Weak Learn algorithm repeatedly in a series of cycle. After the first iteration, new training subset C_1 is obtained. The initial weak classifier $f_1(x)$ is achieved by learning the corresponding training subset C_1. In boosting learning, each example is associated with a weight, and the weights are updated dynamically using a multiplicative rule according to the errors in previous learning so that more emphasis is placed on those examples which are erroneously classified by the weak classifiers learned previously. After *t* iterations, we can obtain *t* weak classifiers .The stronger classifier is a linear combination of t weak classifiers. In our method, we choose SVM as weak classifiers.

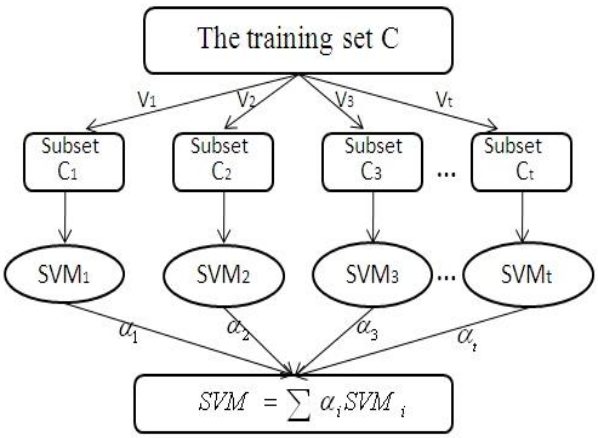

Figure 4. Schematic Diagram of AdaBoost Algorithm

3.2. Weak Classifiers Using SVM

As a data classification of statistical learning method, SVM [18] has intuitive geometric interpretation and good generalization ability, so it has recently gained

popularity within visual pattern recognition. According to the theory, SVM is developed from the theory of Structural Risk Minimization, as shown in Equation 1. For a given sample set $(x_1, y_1),(x_2, y_2),\cdots(x_m, y_m)$ into two classes, where $x_i \in R^N$ is a feature vector and $y_i \in \{-1,\ 1\}$ its class label. Define the hyper plane $(w, \phi(x)) - b = 0$ to make a compromise between class interval and classification errors when the sample is linear inseparable. $(\phi(x_i), \phi(x_j)) = k(x_i, x_j)$ is kernel function, k is corresponding to the dot product in the feature space, transformation ϕ implicitly maps the input vectors into a high-dimensional feature space.

The optimal values for w, b can be found by solving the following minimization problem:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{R}\varepsilon_i \qquad (1)$$

$$y_i(w,\ \phi(x_i) - b) \geq 1 - \varepsilon_i, 0 \leq \varepsilon_i \leq 1$$

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:

$$\max_a \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j (\phi(x_i), \phi(x_j))$$

$$C \geq a_i \geq 0, i = 1,\cdots, n \qquad (2)$$

$$\sum_{i=1}^{n} a_i y_i = 0$$

Where $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^T$ is a Lagrange multiplier which corresponds to $y_i(w,\ \phi(x_i) - b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0$. This paper selected kernel function $k(x_i, x_j) = \exp(-\frac{(x_i - x_j)^2}{2\sigma^2})$

and put it into the Equation 2 to get the final decision function:

$$y(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i y_i k(x_i, x) - b) \qquad (3)$$

The inference problem is to find the best action label *y* for a test video *x*. The weak classifier in this paper is different from establishing SVMs between one against the rest types, instead, adopting the method of one against one to establish a SVM between any two categories. The current sample belongs to which category determined by decision function, and its final type is decided to the category with highest vote.

4. Experiment and Results Analysis

AdaBoost-SVM combined with action descriptors in terms of interest points and FEA define one novel method for action recognition. In this section we evaluate the method for recognizing actions and compare the performance to other approaches.

4.1. Dataset

To test our proposed approach for action recognition, we choose the standard KTH dataset, for evaluating action recognition algorithms. As shown in Figureure5, 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 down sampled to the spatial resolution of 160 × 120 pixels. The examples of the above four scenarios is shown in Figure 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.

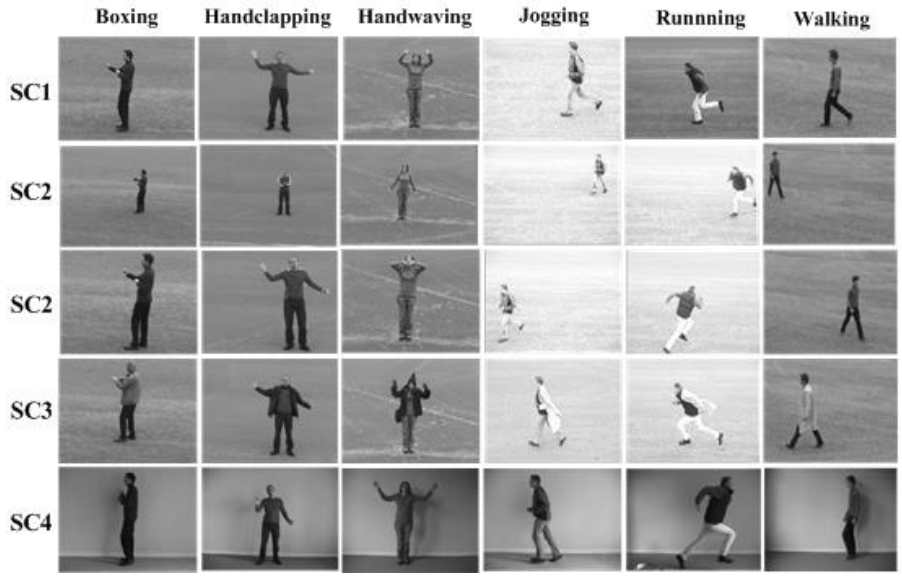

Figure 5. The Examples of KTH Dataset

4.2. Experimental Results

In this part the action recognition performance of the AdaBoost-SVM is assessed by using combined feature of 3D SIFT and FEA on four scenarios (SC1, SC2, SC3 and SC4) according to KTH dataset. 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. Comparison is also made with the performance of the single SVM classifiers, as shown in Table 1 and Figure 6.

Table 1. Testing Results in Portion Scenario

Scenario	SVM	AdaBoost-SVM
SC1	0.9600	0.9933
SC2	0.9200	0.9667
SC3	0.9167	0.9792
SC4	0.9600	1.000

Compared with the recognition rate in Table 1, the recognition accuracy is greatly improved with the application of AdaBoost-SVM classifiers. SC1 and SC4 are more stable than the other two scenarios. We obtained almost 100% correct recognition rate in those two scenarios. Although SC2 and SC3 are relatively complex, the proposed AdaBoost-SVM method can respectively increase the recognition rate by 4% and 6% than traditional SVM method. Compared with the recognition rate in Figure 5, for any action class, the performance of the proposed method is significantly better than the traditional SVM algorithm.

The confusion matrix in each scenario is shown in Figure6. Most of the actions can be correctly recognized. The easily confused actions is "walking" and "running". Confusion between "walking" and "running" can partly be explained by high similarities of these classes.

Overall, the proposed Adaboost-SVM and improved spatio-temporal features achieved the desired recognition results.

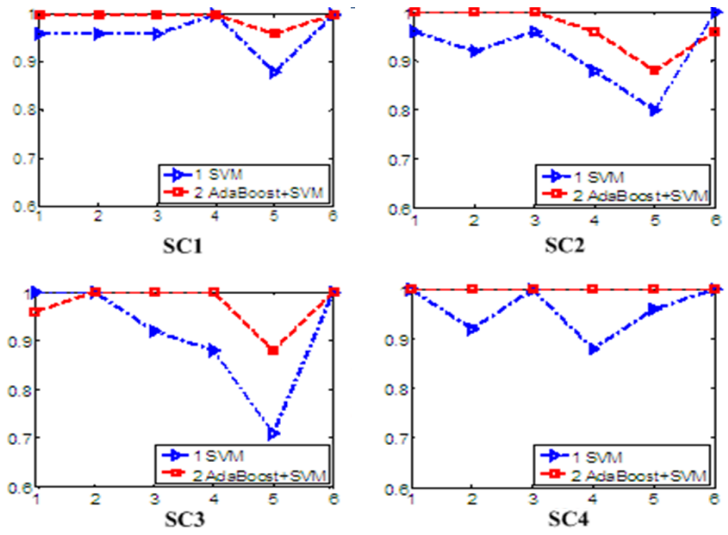

Figure 5. The Comparison Graph of Action Recognition Rate

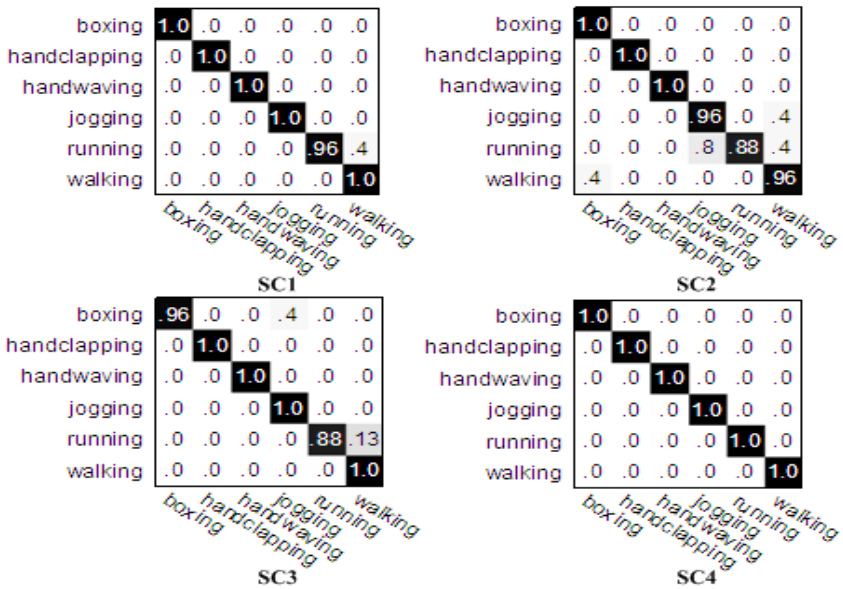

Figure 6. Confusion Matrix in Each Scenario

The comparisons of performance between the proposed method and the recent related Works based on KTH dataset are shown in Table 2. These works are all related to the local spatio-temporal feature. It is worth noting that our method outperforms all of other states of art methods.

Table 2. Comparison with Related Work in Recent Years

Literature	Method	Accuracy
Niebles[8]	3D SIFT BOW +pLSA	83.33%
Klaser *et al.*[20]	3D Gradients +SVM	91.4%
Bregonzie [13]	Interest point clouds +NNC	93.17%
Umakanthan[21]	HOG3D+SVM	92.7%
Our approach	3D SIFT+FEA+AdaBoost-SVM	98.48%

5. Conclusion

This paper proposed a novel action recognition method by using improved spatio-temporal features and Adaboost-SVM classifiers. To obtain more complete representation of the human action, we combine the distribution information of interest points with 3D SIFT descriptor. In order to improve the recognition accuracy, the AdaBoost-SVM is utilized to recognition the human action. The experimental results show that the algorithm based on the AdaBoost-SVM can achieve the accuracy of 98.48% and better performance compared with traditional SVM algorithm.

Acknowledgements

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.

References

[1] D. Weinland, R. Ronfard, E. Boyer, "A Survey of Vision-based Methods for Action Representation", Segmentation and Recognition, Computer Vision and Image Understanding, vol. 2, no. 115, **(2011)**, pp.224–241.
[2] 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. 1, no. 40, **(2010)**, pp. 13–24.
[3] R. Roppe, "A Survey on Vision-based Human Action Recognition", Image and Vision Computing, vol. 28, **(2010)**, pp. 976-990.
[4] X. Li, "Hmm Based Action Recognition Using Oriented Histograms of Optical Flow Field", Electronics Letters, vol. 10, no. 43, **(2007)**, pp. 560–561.
[5] S. Ali and M. Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, no. 32, **(2010)**, pp. 288–303.
[6] X. Cao, B. Ning, P. Yan and X. Li, "Selecting Key Poses on Manifold for Pairwise Action Recognition", IEEE Transactions on Industrial Informatics, vol. 1, no. 8, **(2012)**, pp. 168–177.
[7] A. A. Chaaraoui, P. C. Perez and F. F. Revuelta, "Silhouette-based Human Action Recognition Using Sequences of Key Poses", Pattern Recognition Letters, vol. 15, no. 34, **(2013)**, pp. 1799-1807.
[8] J. C. Niebles, H. Wang and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-temporal Words", International Journal of Computer Vision, vol. 3, no. 79, **(2008)**, 299–318.
[9] J. Zhu, J. Qi and X. Kong, "An Improved Method of Action Recognition Based on Sparse Spatio-temporal Features", Artificial Intelligence: Methodology, Systems, and Applications. Springer, **(2012)**, pp. 240–245.

[10] P. Liu, J. Wang, M. She and H. Liu, "Human Action Recognition Based on 3D Sift and Lda Model" Proceedings of 2011 IEEE Workshop on Robotic Intelligence In Informational Structured Space, **(2011)**, pp. 12–17.
[11] X. Jiang, T. Sun, B. Feng and C. Jiang, "A Space-time Surf Descriptor and its Application to Action Recognition with Video Words", Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery, vol. 3, **(2011)**, pp. 1911–1915.
[12] F. Li, C. Xiamen and J. Du, "Local Spatio-temporal Interest Point Detection for Human Action Recognition", Proceedings of the 5th International Conference on Advanced Computational Intelligence, **(2012)**, pp. 1–10.
[13] M. Bregonzio, S. Gong and T. Xiang, "Recognizing Action as Clouds of Space-time Interest Points", Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, **(2009)**, pp. 1948-1955.
[14] X. Ji, Q. Wu, Z. Ju, Y. W, "Study of Human Action Recognition based on Improved Spatio-temporal features", **(2014)**, vol. 11, no. 5, pp. 500-509.
[15] P. Dollar, V. Rabaud, G. Cottrell and S. Belongi, "Behavior Recognition via Sparse Spatio-temporal Features", Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, **(2005)**, pp.65-72.
[16] 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)**, pp. 357–360.
[17] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions", Machine Learning, vol. 3, no. 37, **(1999)**, pp.297–336.
[18] W. Cheng and D. Jhan, "A cascade classifier using Adaboost algorithm and support vector machine for pedestrian detection", Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, **(2011)**, pp. 1430-1435.
[19] C. C. Chang and C. J. Lin, "Libsvm: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, **(2011)**, pp.1–39
[20] A. Kläser, M. Marszalek, C. Schmid, "A spatio-temporal descriptor based on 3D-gradients. Proceedings of British Machine Vision Conference, **(2008)**, pp.1-8.
[21] S. Umakanthan, S. Denman, S. Sridharan, *et al.*, "Spatio temporal feature evaluation for action recognition", Proceedings of International Conference on Digital Image Computing Techniques and Applications, (**2012)**, pp. 1-8.

Authors

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. Email: jixiaofei7804@126.com

Lu Zhou, is currently a graduate student studying for Master degree in the School of Automation, Shenyang Aerospace University. Her research is focus on the human action modeling and recognition. She has published 3 research papers in this research direction.

Qianqian Wu, received her B.Eng. degree from Lang fang 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.