International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.1 (2016), pp.323-334
http://dx.doi.org/10.14257/ijsip.2016.9.1.31

Multiple Feature Voting based Human Interaction Recognition

Xiaofei Ji, Changhui Wang, Xinmeng Zuo, Yangyang Wang

Shenyang Aerospace University, School of Automation
E-mail:jixiaofei7804@126.com

Abstract

Most of currently interaction recognition methods always need to segment the spatio-temporal features to the individuals involved in the interaction or need to build complex action models to present the human interaction. A novel method is proposed without considering the feature segmentation and complex action model in this paper. The proposed method utilizes two simple features i.e., improved BoW descriptor of interest points and HoG descriptor to respectively represent the local characteristics and global characteristics of human interactions. The classification voting histogram of BoW features and HoG characteristics are obtained by frame to frame nearest neighbor classifier respectively. Finally, recognition result is achieved by weighted fusing the classification voting histogram of these two feature. The method is tested on UT-Interaction dataset. Experiment result show that the method achieved the better recognition performance with simple implementation.

Keywords*: interaction recognition, spatio-temporal interest points, HoG*

1. Introduction

Human interactive behavior recognition and understanding is an important research topic in computer vision community. The related techniques are widely used in the field of intelligent surveillance, human-computer interface, video content-based video retrieval, *etc.* [1] In fact, interactive behavior is a typical human activity in real-world, such as hand shaking, hugging, fighting *etc.* Though many approaches have been proposed to deal with interaction recognition[2], it is still a challenging task due to its large intra-variations, clutter and occlusion, viewpoint changes[3] and other fundamental difficulties. Recent research on interaction recognition can be characterized by two classes of methods:

(1) Interaction Recognition Based on Motion Co-Occurrence: This kind of method considers that the interaction between individuals are composed of a set of temporally ordered elementary actions performed by the different persons involved in the interaction[4-6] . Vahdat *et. al.*, represented the individuals in interactions by a set of key pose, then interaction can be recognized by capturing their spatial and temporal relationship[7]. SLIMANI *et. al*., proposed a co-occurrence of visual words method for human interaction recognition[8]. The method represents the interaction between persons by calculating the number of times visual words occur simultaneously for each person involved in the interaction. The implementation of this method is very simple, however the co-occurrence relationships are not expressive enough to effectively deal with interactions with large variations. Kong *et. al.*, proposed a novel approach by using interactive phrases to describe motion relationships between interacting people. A discriminative model is proposed to encode interactive phrases based on the latent SVM formulation[9]. The method obtains better recognition accuracy, however the training process is relatively complex. In general, this kind of method can achieve more accurate and robust results by exploiting rich contexture information in human interactions.

ISSN: 2005-4254 IJSIP

However the recognition results always depend on the accurate feature segmentation of individual and the stability of individual behavior model.

(2) Interaction is Recognized as a General Action: This kind of method usually represents the interaction as an integral descriptor including all the people involved in the interaction. Then a classier is utilized to classify human interactions[10]. Burghouts *et. al.*, improved the spatio-temporal representation by introducing spatio-temporal layout of actions and obtained successful human interaction recognition[11]. Peng *et. al.*, utilized dense trajectory with four advanced feature encoding methods to achieve human interaction recognition [12]. Li *et. al.*, proposed a hybrid framework which incorporates both global feature and local features to recognize human interactions. The method achieves promising results by respectively using GA based random forest and calculating S-T correlation score as recognition method [13] . This kind of method treats people as a single entity and do not extract the motion of each person from the group. So they do not need segment the feature of individual in the interaction. However the better performance always needs comprehensive motion features.

How to extract discriminative and simple features to describe interactions and design effective recognition methods to fuse different types of features has become an important solution for interaction recognition. The most of above methods [8-9, 13] choose spatio-temporal interest points as fundamental feature to construct the motion feature for interaction recognition due to their simplicity, effectiveness and robustness to cluttered backgrounds. However interest point always model an action as a bag of independent and order-less visual words without considering the spatio-temporal contextual information of interest points. In order to deal with this limitation, a novel feature is proposed by combining local and global information to represent the human interaction. The method do not need segment the features extracted from the image sequence of interaction behavior. The BoW (Bag of Word) descriptor of spatio-temporal interest points in shot length-based video is utilized to represent the local characteristics of human interactive behavior. And the HoG (histogram of gradient) descriptor represents the global characteristics of interaction behavior. Then the recognition result was obtained by fusing the probability of these two kinds of features. The experiments on the UT-Interaction Data set show that the proposed method is very simple and effective.

The rest of this paper is organized as follows. In Section 2, the framework of the proposed method is introduced. Section 3 and Section 4 respectively provides a detailed explanation of local and global feature extraction and representation. And Section 5 gives experimental results and analysis. Finally, Section 6 concludes the paper.

2. The Overview of the Proposed Method

The framework of the approach is shown in Figure 1. The framework includes the following modules:

(1) Interaction Detection: In order to improve recognition accuracy and efficiency, interaction detection in video is performed before feature extraction. The foreground information of the interaction is obtained by using frame difference. Redundant information in the frame can be eliminated by using the frame difference. The interaction process is shown in Figure 2.

(2) Feature extraction: Improved BoW descriptor of spatio-temporal interest points is used to represent local characteristics of interactions in this paper. It is robust to the changes in viewpoint and environment. Simultaneously HoG feature is used to represent the global characteristics.

(3) Voting classification: Classification voting histograms of these two features are

obtained by the nearest neighbor classifier. The nearest neighbor technique is very simple, highly efficient and effective in the field of pattern recognition.

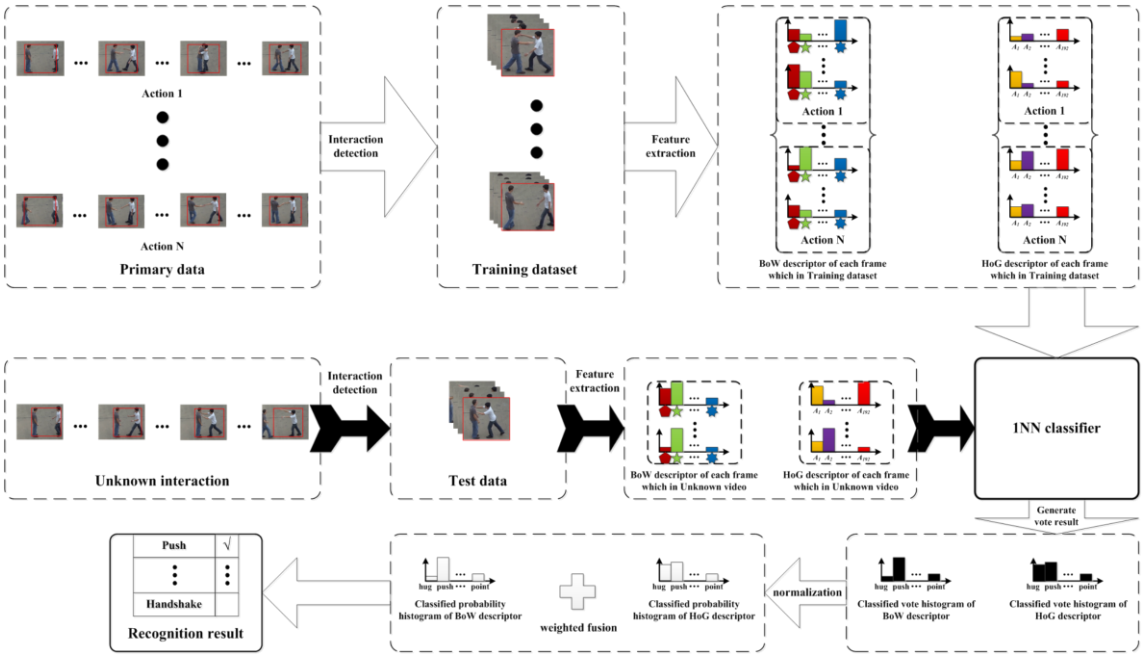

Figure 1. The Interaction Recognition Framework of the Proposed Approach

(4)Weighted fusion of voting histograms: The classification probability histogram are obtained by normalize these classification voting histograms of two features separately. Finally recognition results are obtained by weighted fusing the classification probability histogram of two features.

3. Local Feature Extraction and Representation

3.1. Interest Points Detection

In computer vision, spatio-temporal interest point was always used to represent the local characteristics of the human behavior in the image sequences. The most widely used detection method of interest points was proposed by Dollars[14]. The method calculates function response values based on the combination of Gabor filter and Gaussian filter. The extreme values of local response can be considered as spatio-temporal interest point of the interaction behavior in the image sequence. The response function is given as:

$$r = (I \times g(x, y; \sigma) \times h_{ev}(t, \tau; w))^2 + (I \times g(x, y; \sigma) \times h_{od}(t, \tau; w))^2 \quad (1)$$

Where $g(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{(x^2+y^2)/2\sigma^2}$ is the Gaussian smoothing kernel of the Gaussian filter. $h_{ev}(t, \tau; w) = -cos(2\pi wt)e^{-t^2/\tau^2}$ and $h_{od}(t, \tau; w) = -\sin(2\pi wt)\, e^{-t^2/\tau^2}$ are the 1D Gabor filter applied temporally, σ and τ are spatial and temporal scale in detection respectively.

Dollar's method is sensitive to both background noise and highly textured object foreground areas regardless their relevance to capturing the dynamics of action observed. To overcome this problem, a different interest points detection method is utilized on the interaction video [15]. Firstly, the region of interest is detected by using frame difference. Then Gabor filtering is performed on the detected regions of interest from different orientations.

3.2. The 3D-SIFT Descriptor of Spatio-Temporal Interest Points

3D SIFT descriptor is chosen to represent Spatio-temporal Interest Point. There are three steps to generate the descriptors:

(1) Extract the spatio-temporal cube from the image sequence and divide it into fixed size unit sub-cubes.

(2) Calculate the spatio-temporal gradient histogram of each unit cube by using faceted sphere.

(3) The 3D SIFT descriptor of spatio-temporal interest point is formed by combining all the unit cube histograms [16]. In this paper, the 12× 12 × 12 pixel size cube is divided into 2 sub cubes. In our previous paper about recognition single human action [17], we adopts 32 faceted at 32 gradient directions for descriptor, so the whole features of each point are 32 × 8 dimensions. Those parameters are used to describe spatio-temporal interest point of interactive behavior in this paper.

3.3. The BoW Descriptor of Spatio-Temporal Interest Point

The conventional BoW descriptor of spatio-temporal interest point is usually extracted throughout the whole video, and the spatio-temporal contextual information of interest points always be ignored. The BoW descriptor of interest points in shot length-based video is utilized to represent the local information of human interaction[18]. The graphical representation of BoW Descriptor Generation is shown in Figure 2.

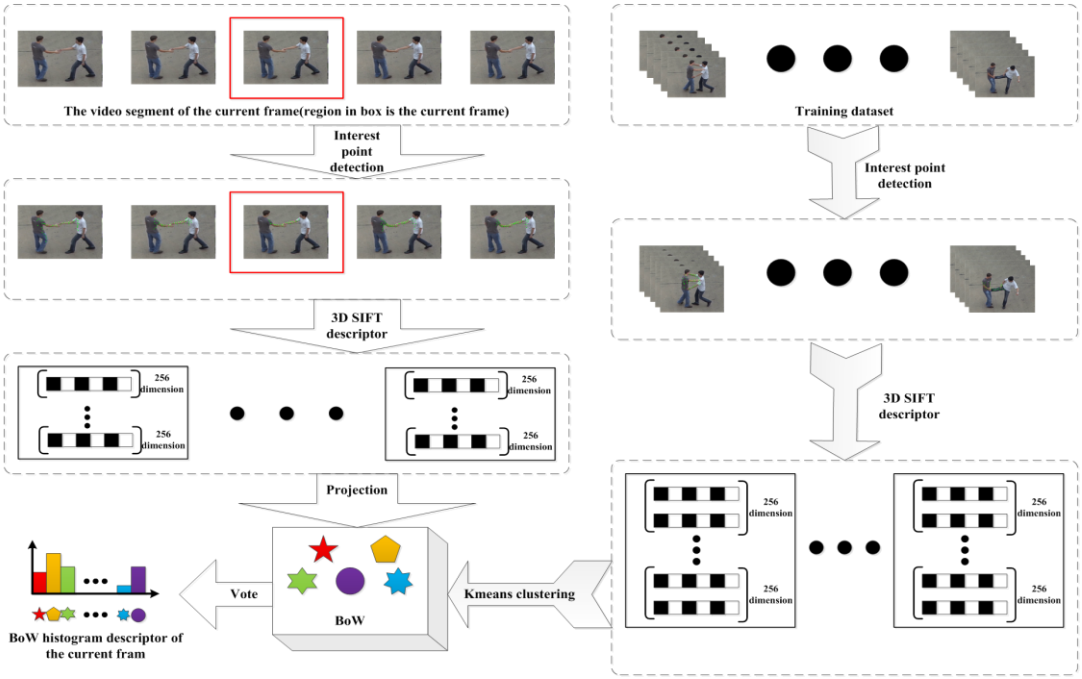

Figure 2. The Graphical Representation of BoW Descriptor Generation

The detail is following:

(1) The 3D SIFT descriptors areclustered into a C-dimensions vocabulary after the 3D SIFT description of spatio-temporal interest point are calculated in single human action video.

(2) When new frame of test video is input, the 3D SIFT descriptor of

spatio-temporal is extracted in the neighbor F frames .

(3) Then the descriptors are projected into the C-dimensions vocabulary by minimizing the Euclidean distance between them and the vocabulary words. The frequency of the vocabulary words were counted as the BoW representation of interest points for the current frame. The best performance with C=60 and F=4 were found in our previous paper.

4. The Global Feature Extraction and Representation

In order to perform recognition at a faster speed, it is necessary to extract a small amount of raw feature data with a simple and discriminative feature representation. It has been proved that grid based HoG descriptors significantly outperform existing feature sets for human detection in a previous study [19]. HoG feature reflects the edge gradient information of human motion, do not need complex edge detection process. This method can overcome the disturbance changes due to illumination, scale, wearing and background, even in complex background environment still has strong stability. The HOG representation is formed by calculating the gradient histogram in local areas and statistical image under the main idea that the local target appearance and shape can be described by the density distribution of light intensity gradient or edge direction in an image.

The extraction process of HoG feature requires two steps:

(1) The image is divided into a plurality of connected non overlapping cells with the same size;

(2) Histogram of oriented gradient of each pixel is statistics in each cell, then all histograms of all cell unites are combined to generate the final descriptor.

The HOG features are extracted to represent global characteristics of interactive behavior in this paper. The HOG descriptor for an interest region of interactive behavior is described as a feature vector $D_s = (s_1,...,s_{n_s}) \subseteq R^{n_s}$ by dividing the action interest region into $n_s = 4$ square grids $R_1,...,R_{n_s}$. Histogram of oriented on 12bins can be utilized to encode each sub-region, as shown in Figure 4 (b)(e). Then all the histograms can be concatenated to form a 4 × 4 × 12 raw feature vector.

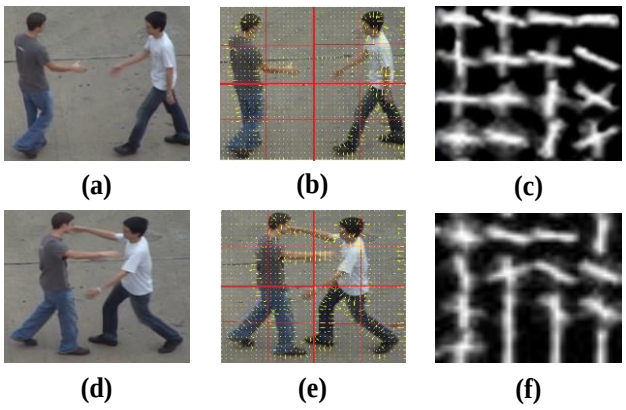

Figure 4. The Graphical Representation of HoG Descriptor Generation

5. Classifier Design

5.1. Frame Nearest Neighbor Recognition Method

The nearest neighbor classify algorithm is not only a simple and effective identification

method, but also has a fast recognition speed to recognize single human action [20]. In order to realize the real-time detection, the nearest neighbor classifier is chosen to recognize the extracted features respectively. The method is shown as follows:

Supposing that there are c classes as $w_1,w_2,...,w_c$, each class has the number of N_i marked sample, then the discriminant function of class w_i is shown as Eq.2:

$$g_i(x) = \min \left\| x - x_i^k \right\|, k = 1,2,...,c \tag{2}$$

The subscript i of x_i^k means class w_i and k is the kth sample among total N_i in classes w_i. According to the function above, the decision rule can be defined as Eq. 3:

$$If \qquad g_j(x) = \min g_i(x), i = 1,2,...,c \tag{3}$$

$$So \qquad x \subseteq w_j$$

This decision method is called nearest neighbor method. The calculation formula of Euclidean distance between samples is shown as Eq. 4:

$$D = \sqrt{\sum_{i=1}^{N} (A_i - B_i)^2} \tag{4}$$

A and B are feature vectors and N is the number of the feature vectors.

The recognition method used in this paper is also called frame to frame nearest neighbor. Firstly the training samples with known category have been chosen to form training set. The frames included the same interaction category have the same symbol in the training set. Then input the test sequences, the classifier try to forecast the symbol of test interaction sequences to be one action type by respectively calculating the Euclidean distance between feature of each test frames and each training frames feature. Then vote for the action category which the frame with minimum distance in the training samples belongs to. Finally the category of test sequence will be recognized as the action class symbol which has the highest votes.

5.2. Two Kinds of Feature Decision Fusion

The classification voting histogram of BoW descriptor and HoG descriptor of the test video sequences are generated by using the nearest neighbor classifier. Then, the classification voting histogram of BoW descriptor and HoG descriptor are normalized to generate the classification probability histogram of BoW descriptor and HoG descriptor. Finally, the recognition result is obtained by weighted combining the classification probability histogram of BoW descriptor and HoG descriptor. The decision fusion process is shown as Eq.5:

$$V = v_1, v_2, ..., v_c \tag{5}$$

Vector V represents the classification voting histogram of BoW descriptor or HoG descriptor of the unknown video sequence. c represents the number of interaction classes which the training set contains. The classification probability histogram P is generated by normalizing the classification vote histogram V. The calculation formula is shown as Eq.6, pi represents the probability of the current test video to each class:

$$p_i = \frac{v_i}{\sum_{i=1}^{c} v_i}, i = 1,2,...,c \tag{6}$$

Then, these two kinds of classification probability histogram are weighted fused by using the Eq.7:

$$P_f = w_{BoW} \times P_{BoW} + w_{HoG} \times P_{HoG} \tag{7}$$

The initial weights are obtained by experiment. Testing results of the experiment are shown in Figure 5. The best result was achieved when the weight of BoW is 66% and the weight of HoG is 34%.

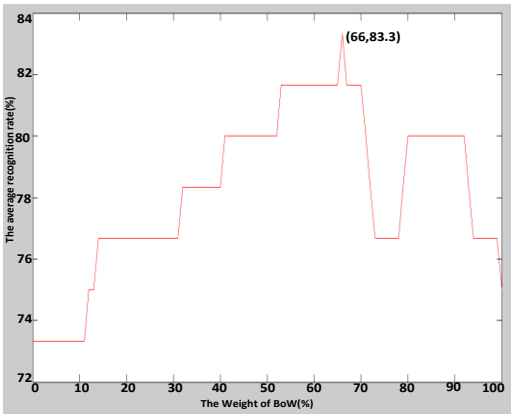

Figure 5. The Weight Testing Result **Figure 6. The Examples of the Dataset**

6. Algorithm Verification and Results Analysis

6.1. Dataset

To test the effectiveness of our approach, the UT-Interaction set 1 benchmark dataset[21] was chosen, which contains 6 classes of human interactive behaviors performed by 15 peoples: shake-hands, hug, kick, point, punch, push. Each class contains 10 video sequences. Some challenging factors in this dataset include moving background, cluttered scenes, camera jitters/zooms and different clothes. The segmented UT-interaction sequences were used for evaluating the recognition accuracy and speed of our method in the experiments. As presented in Figure 6, there are 6 action types in the dataset.

6.2. Testing Results

Recognition experiment are performed by using combined the classification voting histogram of BoW descriptor and HoG descriptor on UT-Interaction set 1 in this part. Leave-one-out cross validation method is adopted throughout the process. In turns using one action of each action class as test samples, and the remaining others as the training set, circulation continue until all actions are completed testing. The experimental results are shown in Table 1, and the confusion matrixes are shown in Figure 7.

Table 1. The Recognition Results (%) of Proposed Method

Feature	HoG	BoW	HoG+BoW
Hand shake	80	100	90
Hug	90	90	100
Kick	80	70	80
Point	80	80	90
Punch	60	20	60
Push	50	90	80
Average	73.3	75	83.3

We can find that the better recognition result has obtained on actions ‘hug’ ‘kick’ ‘point’ ‘punch’ by using weighted fusion of classification probability histogram of BoW feature and HoG feature. For action ’hug’ and ‘point’, the recognition accuracy(100% , 90%) of combined features (BoW+HoG) has increased by 10% than BoW features and HoG features using alone. For action ‘kick’ and ‘punch’, the same recognition rate(80% , 60%) was obtained by using combined features (HoG+BoW) and HoG, but it has raised(10% ,40%) than BoW used alone. Although , for action ‘hand shake’ and ‘push’, the recognition rat(90% , 80%) of combined features (BoW+HoG) is lower than the recognition rat (100% , 90%) by BoW used alone , but it has raised (10% , 30%) than HoG used alone.

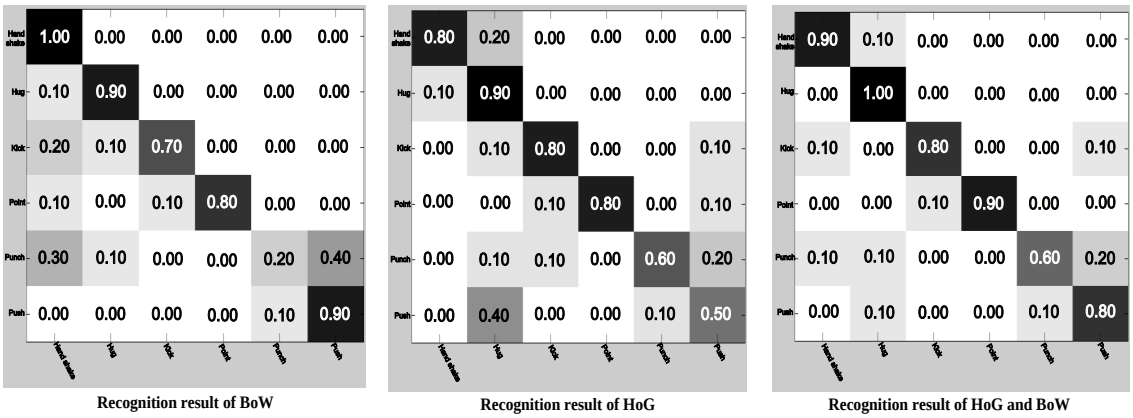

Figure 6. The Confusion Matrix of the Proposed Method

6.3. The Comparison of the Performance

The comparisons of performance between the proposed method and the recent related works based on UT-Interaction dataset are shown in Table 2.

Table 2. Comparison with Related Work in Recent Years

Literature	Year	Method	Accuracy
Ryoo *et al*.[21]	2009	Spatio-temporal interest point+relationship match kernel	70.8%
Mukherjee *et al*.[22]	2011	Bipartite graph+key pose doublets	79.17%
Brendel *et al*.[23]	2011	2D+t tubes+spatio-temporal relationships graph model	78.9%
Ryoo[24]	2011	3D spatio-temporal Cuboid+ dynamic BoWs	71.7%
Kong *et al*.[25]	2014	global template +local 3D feature+	85%

		discriminative model	
SLIMANI *et al*.[8]	2014	3D XYT spatio-temporal volume + BoW + co-occurrence matrix	41%
Our approach	2015	HoG+BoW+1NN	83.33%

Obviously, our approach has achieved a good recognition result. The recognition rate of [25] are slightly higher than ours, however the method need complicated feature extraction and require a plurality of complex recognition model. Our approach need not segment the feature of interactive behavior to individuals and create any complex discriminant model. So our method outperforms all of other state of methods.

7. Conclusion

This paper proposed a novel approach by using two simple features *i.e*., improved BoW descriptor of interest points and HoG descriptor to represent global characteristics and local characteristics of interactions. Classification voting histogram of BoW and HoG features are obtained by using frame to frame nearest neighbor classifier respectively. Finally, recognition results obtained by weighted fusing them. Compared with previous research, neither our method requires assign the features to the individuals in interaction, nor build the complex model for recognition. The proposed classifier has a simple structure with better adaptability. Experimental results show that the proposed combined feature can effectively compensate for performance when the two kinds of feature are used alone. Compared with the results of other studies, the proposed method is simple and effective. In the future, we will exploit rich contexture information in human interactions to help achieve more accurate and robust results.

Acknowledgments

The project supported by the Program for Liaoning Excellent Talents in University (No. LJQ2014018) and the Scientific Research General Project of Education Department of Liaoning Province, China (No. L2014066)

References

[1] D. Weinland, R. Ronfard and E. Boyer, "A survey of vision-based methods for action representation, Segmentation and Recognition", Computer Vision and Image Understanding, vol. 2, no. 115 **(2011)**, pp. 224–241.
[2] Y. Cao and D. Barrett, "Recognizing human activities from partially observed videos", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, USA, **(2011)**, pp. 2658-2665.
[3] X. Ji and H. Liu, "Advances in View-invariant human motion analysis: a review", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 1, no. 40, **(2010)**, pp. 13–24.
[4] L. Meng and L. Qing, "Activity recognition based on semantic spatial relation", Proceedings of the 21 International Conference on Pattern Recognition, Tsukuba, Japan, **(2012)**, pp. 609-612.
[5] A. Patron-Perez, M. Marszalek, I. Reid and A. Zis-serman, "Structured learning of human interactions in TV shows", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 34, **(2012)**, pp. 2441-2453.
[6] M. Raptis and L. Sigal, "Poselet Key-framing: A model for human activity recognition", Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, **(2013)**, pp. 2650-2657
[7] A. Vahdat, B. Gao, M. Ranjbar and G. Mori, "A discriminative key pose sequence model for recognizing human interactions", Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, **(2011)**, pp. 1729-1736
[8] K. SLIMANI, Y. BENEZETH and F. SOUAMI, "Human interaction recognition based on the co-occurrence of visual words", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, USA, **(2014)**, pp. 461-466.

[9] Y. Kong, Y. Jia and Y. Fu, "Interactive Phrases: semantic descriptions for human interaction recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 9, no. 36, **(2014)**, pp. 1775-1788.
[10] T. Yu, T. Kim and R. Cipolla, "Real-time action recognition by spatio-temporal semantic and structural forests", Proceedings of the 21st British Machine Vision Conference. Aberystwyth, United kingdom, **(2010)**, pp. 1-12.
[11] G. J. Burghouts and K. Schutte, "Spatio-temporal layout of human actions for improved bag-of-words action detection", Pattern Recognition Letters, vol. 34, **(2013)**, pp. 1861-1869.
[12] X. Peng, Q. Peng and Y. Qiao, "Exploring dense trajectory feature and encoding methods for human interaction recognition", Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, Huang shan, China, **(2013)**, pp. 23-27.
[13] N. Li, X. Cheng, H Guo and Z Wu, "A Hybrid method for human interaction recognition using spatio-temporal interest points", Proceedings of the 22nd International Conference on Pattern Recognition, Stockholm, Sweden, **(2014)**, pp. 2513-2518.
[14] P. Dollar, V. Rabaud, G. Cottell and S. Belongie, "Behavior recognition via sparse spatio-temporal features", Proceedings of 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, **(2005)**, pp. 65-72.
[15] M. Bregonzio, S. Gong and T. Xiang, "Recognising action as clouds of space-time interest points", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, **(2009)**, pp. 1948-1955.
[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, ACM, **(2007)**, pp. 357-360.
[17] X. Ji, Q. Wu, Z. Zhao and Y. Wang, "Study of human action recognition based on improved spatio-temprotal features", International Journal of Automation Computing, vol. 5, no. 11, **(2014)** pp. 500-509.
[18] X. Ji, C. Wang and Y. Li, "A view-invariant action recognition based on multi-view space hidden markov models", International Journal of Humanoid Robotics, no. 1, **(2014)**.
[19] N. Dalad and B. Triggs, "Histogram of oriented gradients for human detection", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA, **(2005)**, pp. 886-893.
[20] X. Ji, L. Zhou and Y. Li, "Human Action Recognition Based on AdaBoost Algorithm for Feature Extraction", Proceedings of IEEE International Conference on Computer and Information Technology, **(2014)**, pp. 801-805.
[21] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities", Proceedings of the IEEE International Conference on Computer Vision. Kyoto, **(2009)**, pp. 1593-1600.
[22] S. Mukherjee, S. Biswas and D. Mukherjee, "Recognizing interaction between human performers using "key pose doublet", Proceedings of the ACM Multimedia Conference, Scottsdale, AZ, United states, **(2011)**, pp. 1329-1332.
[23] W. Brendel and S. Todorovic, "Learning spatio-temporal graphs of human activities", Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, **(2011)**, pp.778-785.
[24] M. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos", Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, **(2011)** pp. 1036-1043.
[25] Y. Kong, W. Liang, Z. Dong and Y. Jia, "Recognizing human interaction from videos by a discriminative model", Institution of Engineering and Technology Computer Vision, vol. 4, no. 8, **(2014)**, pp. 277-286

Authors

Xiaofei Ji, he 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 The Program for Liaoning Excellent Talents in University (No. LJQ2014018). E-mail: jixiaofei7804@126.com (Corresponding author)

Changhui Wang, He received his B.Eng.degree in Measurement and control technology and instrument from Yanshan University, Linren college,China,in2013.He is currently a graduate student studying for Master degree in the school of Automation, Shenyang Aerospace University. His research is focus on the human interactive behavior recognition.

Xinmeng Zuo, He received his B.Eng. Degree in Automation Engineering from the Liaoning Shihua University, Fushun, China, in 2014. He is currently a graduate student studying for Master degree in the School of Automation, Shenyang Aerospace University. His research is focus on the two-person interaction detection and recognition.

Yangyang Wang, She 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.