Human motion correction and representation method from motion camera

Hong-Bo Zhang[1,2], Feng Guo[3], Miaohui Zhang[4], Ying Lin[3], Tsung-Chih Hsiao[1]

[1]*Department of Computer Science and Technology, Huaqiao University, Xiamen, People's Republic of China*
[2]*Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen, People's Republic of China*
[3]*School of Information Science and Engineering, Xiamen University, Xiamen, People's Republic of China*
[4]*Institute of Energy, Jiangxi Academy of Sciences, Jiangxi Province, People's Republic of China*
E-mail: betop@xmu.edu.cn

Published in *The Journal of Engineering*; Received on 2nd May 2017; Accepted on 13th June 2017

Abstract: Motion estimation is a basic issue for many computer vision tasks, such as human–computer interaction, motion objection detection and intelligent robot. In many practical scenes, the object movement goes with camera motion. Generally, motion descriptors directly based on optical flow are inaccurate and have low discrimination power. To this end, a novel motion correction method is proposed and a novel motion feature descriptor called the motion difference histogram (MDH) for recognising human action is proposed in this study. Motion estimation results are corrected by background motion estimation and MDH encodes the motion difference between the background and the objects. Experimental results on video shot with camera motion show that the proposed motion correction method is effective and the recognition accuracy of MDH is better than that of the state-of-the-art motion descriptor.

1 Introduction

Motion estimation and recognition is the foundation of many computer vision works, especially for object motion analysis in visible light camera. It is widely used in many applications, such as human–machine interaction, video surveillance, event retrieval and intelligent vehicles. In many practical scenes, the object movement goes with camera motion. So recognising human motion from motion camera is a hot research topic in computer–human interaction [1, 2] and computer vision [3, 4]. The approaches of human action recognition involve motion estimation/representation, object detection and trajectories. In most of these video analysis tasks, the motion feature is popularly used as a low-level vision feature and plays an important role. However, in real scenes, owing to the movement of the camera and objects, error exists in motion estimation, reducing the discrimination power of the motion descriptor.

For motion recognition in complex scenes, especially in a camera motion environment, how to model camera motion is still an open issue. Wang and Schmid [5] estimated camera motion by matching feature points between frames and using the motion boundary histogram (MBH) to represent motion. Unfortunately, there is no clean solution to this problem. Towards this end, we propose a novel correction method for motion estimation results and a novel motion descriptor called the motion difference histogram (MDH) is calculated, which regards the background motion as camera motion.

To estimate motion and compute MDH, the dense optical flow of the video is extracted via the Lucas–Kanade (LK) algorithm [6]. The maximising component is regarded as camera motion, and the real motion is the relative motion between the optical flow and camera motion. Finally, the histogram of the orientation of real motion is computed as MDH.

To verify the accuracy of motion correction and the discrimination power of MDH, we use the conventional bag-of-words (BOW) model to represent the motion. The video is regarded as a set of spatiotemporal interest points (STIPs) detected by a 3D Harris algorithm [7]. MDH is used for the motion representation of STIP, and the visual word vocabulary of the action is constructed. Finally, the motion is regarded as the visual word feature, and the support vector machine (SVM) classifier is trained for motion recognition. Fig. 1 shows the motion recognition strategy of BOW model. In this work, we focus on motion estimation and representation, the BOW model is a simple pattern recognition model to evaluate the motion descriptor.

The contributions of our work are threefold:

(i) We propose a specific approach to estimate background/camera motion and the human motion is corrected by the difference of camera motion and optical flow.
(ii) We propose a novel motion descriptor for discriminative action representation. The experimental results show that the discrimination of MDH is better than that of the state-of-the-art motion descriptor.
(iii) The proposed motion descriptor is sufficiently general for other off-the-shelf vision tasks. We are open to more robustness model to replace BOW motion recognition model. Currently, we place greater emphasis on the accuracy of motion estimation and motion representation.

The remaining of this paper is organised as follows. Section 2 reviews some related works. Section 3 describes the proposed method. Section 4 presents and discusses our experimental results. Finally, Section 5 concludes the paper.

2 Related works

The mainly approaches of motion estimation from camera involve in optical flow, frame/background difference method and object tracking. The optical flow method [8] tries to calculate the motion between two frames based on the optical flow constraint equation which assume the motion remain the same in very short time. Frame difference method needs a good and robustness background model. Moreover, object tracking is based on accuracy object detector and tracker. However, due to camera motion, the motion estimation of these methods is inaccuracy. In this work, we propose a new method correction method to calculate real motion from the motion estimation of optical flow.

To verify the accuracy of the proposed motion correction method, motion descriptor based on motion correction results is calculated for STIP to recognition human motion. Many studies in the literature indicate that STIP is widely used in human action

doi: 10.1049/joe.2017.0159

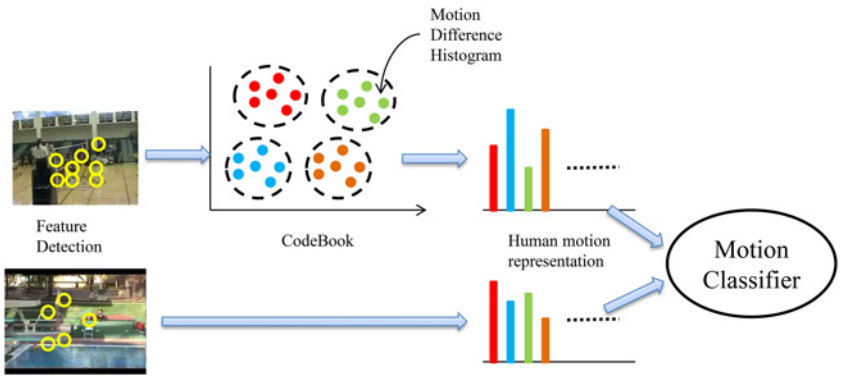

Fig. 1 *Framework of the BOW model*

recognition tasks owing to its robustness and good performance. In this study, we also focus on STIP and discuss the motion descriptor of STIP. Generally, two descriptor types are used to represent motion: absolute motion descriptor and relative motion descriptor. The absolute motion descriptor is computed directly based on optical flow, such as the histogram of the orientation of optical flow (HOF) [9]. This approach is simple but inaccurate owing to background motion, especially camera motion. The relative motion descriptor receives more attention because of its good performance in human action recognition. Frequently used relative motion descriptors include MBH [5] and Internal Motion Histograms (IMHcd).

In this study, we also discuss the relative motion descriptor and propose a novel descriptor named MDH. In contrast to these descriptors, MDH estimates the camera motion by maximising the statistical distribution of the optical flow. The real motion of each pixel is expressed by subtracting from the camera motion.

To verify the discrimination power and effectiveness of MDH, we use the BOW model to construct the action representation based on the motion descriptor. An SVM classifier is constructed to recognise action. In this study, the emphasis is on the effective of MDH, which is indicated by comparing MBH with IMHcd. BOW is wildly used in many vision tasks. Wang *et al.* [10] used the K-means algorithm to create visual words, and action is expressed as the word sequence. Niebles *et al.* [11] used an unsupervised learning algorithm to create a visual word codebook, and actions were recognised via the probabilistic latent semantic analysis (pLSA) or latent Dirichlet allocation (LDA) algorithm. In this study, the emphasis is on the effective of motion correction and MDH, which is indicated by comparing with MBH and IMHcd.

3 Proposed motion correction method and motion descriptor

We describe the proposed motion correction method and motion descriptor for STIP as follows.

3.1 Motion correction method

To calculate precision motion from motion camera, it is a necessity to eliminate the influence of camera motion. Towards this end, we assume the background motion is raised by camera motion and the background motion is argued as camera motion. Thus, relative motion is a good solution. Firstly, the optical flow $\boldsymbol{I}$ is computed based on pyramidal frames structure. With camera motion, the optical flow $\boldsymbol{I}$ is the sum of object motion $\boldsymbol{I}_r$ and camera motion $\boldsymbol{I}_c$

$$\boldsymbol{I} = \boldsymbol{I}_r + \boldsymbol{I}_c \tag{1}$$

where these motion vectors can be decomposed into the horizontal and vertical directions (x and y directions) as follows:

$$\begin{aligned} I_x &= I_{rx} + I_{cx} \\ I_y &= I_{ry} + I_{cy} \end{aligned} \tag{2}$$

where I_{rx} indicates the object motion in the x-direction, I_{ry} indicates the object motion in the y-direction, I_{cx} indicates the camera motion in the x-direction and I_{cy} indicates the camera motion in the y-direction.

In the same image, the camera motion is fixed for all points. The object motion vector is estimated by solving I_{rx} and I_{ry}. The key is how to estimate the camera motion. However, estimation of camera motion directly from video data is still a challenging problem in computer vision. In this work, the background motion is estimated by analysing the optical flow of dense interest points. The background motion is regarded as camera motion to compute object motion.

To compute the background motion, the local interest points of the image are extracted first. In this work, we use the Harris corner points as the detector and extract the optical flow of these interest points via the LK algorithm. Some examples are shown in Figs. 2*a* and *b*.

The optical flow of these points is decomposed into the x and y directions. The value is divided into ten intervals. The distribution of points is then accumulated. Each value in the interval indicates the number of points. Examples are shown in Fig. 2*c*. The maximisation of the histogram is regarded as background motion (camera motion) because the overwhelming majority of the movement points are caused by camera motion, and the movement patterns of these points are consistent. The background motion pattern is shown in Fig. 2*d*. The relative motion can be estimated by using (2).

3.2 Motion descriptor and recognition method

After precise relative motion estimation, to evaluate the effective of motion correction and motion representation, we use the relative motion feature to recognise human motion. A new descriptor named motion difference histogram (MDH) is computed in the spatiotemporal domain of STIP. The domain is divided into 3^*3^*2 cells, and the histogram of the orientation of relative motion is computed in each cell. The angles of $0^\circ - 360^\circ$ are divided into nine intervals. Finally, by combining the histograms of these cells, the dimension of MDH is $3^*3^*2^*9 = 162$. The computational process of MDH is shown in Fig. 3.

To recognise human action, the video is represented as a histogram feature of a visual word dictionary. To create a visual word dictionary, we use the K-means algorithm for each category based on STIP and the motion descriptor. The length of the dictionary in each category is k. Finally, the dictionary is

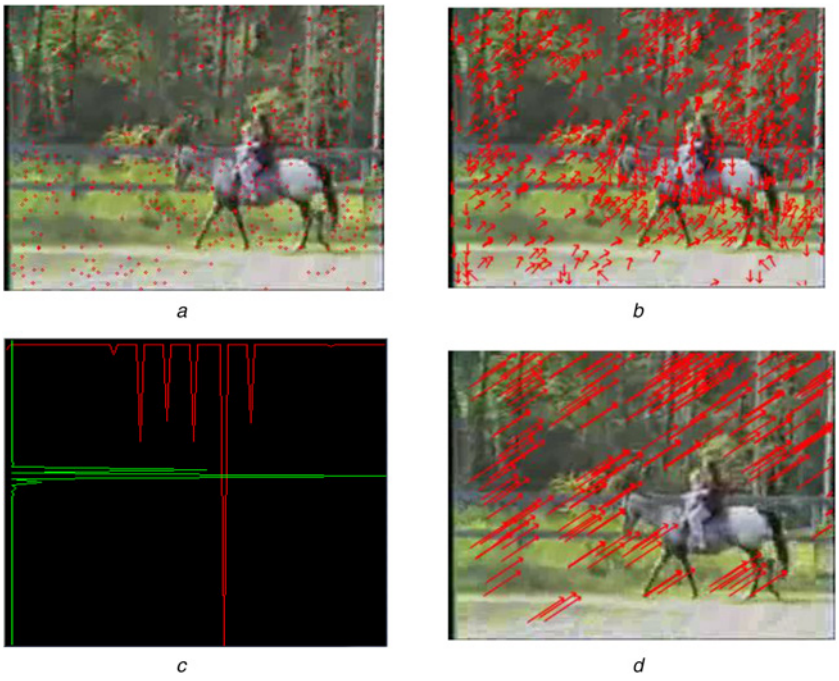

Fig. 2 *Example of the background motion estimation*
a Harris corner points
b Optical flow of Harris corner points
c Distribution of optical flow in the horizontal and vertical directions (red indicates the horizontal direction, and green indicates the vertical direction)
d Background motion

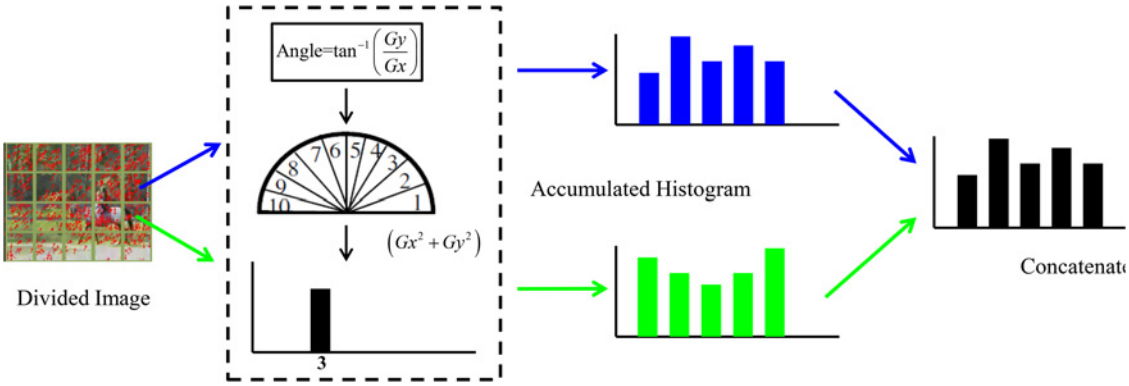

Fig. 3 *Framework of the MDH feature*

Fig. 4 *Examples of YouTube Action dataset*

Table 1 Comparison of different descriptors for action recognition in the YouTube dataset

S	MBH, %	HOF [9], %	IMHcd, %	MDF, %	HOG [9], %	HNF [9], %	HOGNMDH, %
25th	43.33	47.46	**61.67**	55.93	45.76	50.85	42.37
24th	43.28	40.58	**51.47**	50.72	44.93	53.62	**56.52**
23rd	42.37	55.93	**62.71**	57.63	52.54	**69.49**	**64.41**
22nd	23.08	**41.54**	**41.54**	**41.54**	35.83	49.23	**52.31**
21th	60.32	55.56	58.73	**61.90**	50.79	57.14	**68.25**
20th	36.11	**52.78**	51.39	50	48.61	**55.56**	**55.56**
19th	37.50	34.38	**48.44**	31.25	53.13	48.44	**51.56**
18th	35.09	45.61	36.84	**54.39**	**50.88**	43.86	52.63
17th	37.29	52.54	49.15	**64.41**	57.63	**61.02**	59.32
16th	35.94	42.19	**43.75**	37.50	**50**	46.88	**50**
15th	36.67	48.33	48.33	**61.67**	**76.67**	**60**	**65**
14th	33.33	44.44	**46.30**	**46.30**	40.74	33.33	37.04
13th	42.11	31.58	**45.61**	35.09	49.12	**52.63**	50.88
12th	36.21	48.28	48.28	**50**	53.45	**55.17**	51.72
11th	47.76	49.25	**67.16**	62.69	59.70	**61.19**	**62.69**
10th	33.33	66.67	56.67	**58.33**	53.33	61.67	**61.67**
9th	25	**46.88**	**46.88**	42.19	**53.13**	48.44	**60.94**
8th	25.86	**53.45**	31.03	44.83	**58.62**	50	**63.79**
7th	47.06	52.94	52.94	**57.35**	**67.65**	**61.76**	**66.18**
6th	36.51	42.86	**50.79**	**50.79**	44.44	50.79	**52.38**
5th	**47.06**	45.59	44.12	**47.06**	**48.53**	47.06	**55.88**
4th	52.94	**58.82**	39.71	41.18	48.53	**60.29**	55.88
3rd	27.59	**37.93**	36.21	**37.93**	**44.29**	41.38	**44.83**
2nd	35.82	47.76	35.82	**58.21**	55.22	**70.15**	**67.16**
1th	33.33	59.65	47.37	**64.91**	**61.40**	**61.40**	**63.16**
avg	38.20	48.12	48.12	**50.55**	52.20	54.05	**56.49**

Bold values indicate results of the proposed and best results

$V = \{w_{11}, \ldots, w_{1k}, \ldots, w_{c1}, \ldots, w_{ck}\}$, where C indicates the number of action categories. The video is expressed as the C^*k histogram feature by mapping STIP to the dictionary.

After computing the video feature, the SVM classifier is trained for action recognition. In this work, the RBF kernel is used to train and predict the SVM classifier

$$K(H_i, H_j) = \exp\left(-\frac{1}{2\sigma^2}^* \left|H_i - H_j\right|\right) \quad (3)$$

where H_i and H_j are the features of the video (visual word histogram), and σ^2 is estimated by cross-validation.

4 Experimental results

4.1 Dataset and parameter setting

In this study, we discuss the motion correction and motion descriptor method in camera motion scene. The accuracy and effective of the proposed motion correction and descriptor method are verified in human motion recognition challenge. The method is designed based on YouTube dataset [12], which contains 11 actions ($C = 11$): 'basketball shooting', 'biking/cycling', 'diving', 'golf swinging', 'horseback riding', 'soccer juggling', 'swinging', 'tennis swinging', 'trampoline jumping', 'volleyball spiking' and 'walking with a dog'. All of the video in this dataset are collected from the YouTube website. This is a challenge owing to the large variations in camera motion, object appearance, object pose, object scale, viewpoint, background clutter and illumination conditions. Each action has 25 subjects ($S = 25$) containing more than 4 different environments ($E \geq 4$) for a total of 1599 videos. Fig. 4 presents some examples from YouTube dataset.

4.2 Performance evaluation of human motion recognition

To verify the accuracy of motion correction and the discriminative of the proposed descriptor, we compared MDH with MBH, HOF [9] and IMHcd in the YouTube dataset. In our experiment, we used 25-fold leave-one-out cross-validation to measure the performance of the proposed method. In each round, one subject is selected as testing data $N_{\text{test}} = C^*E$, and the remaining are the training data, for a total of $C^*E^*(S-1)$. To create the dictionary, the cluster number is set at $k = 20$. The accuracy is the average of 25 rounds. The comparison result is shown in Table 1.

Table 2 Comparison of different the cluster number for action recognition in YouTube dataset

	$K = 20$, %	$K = 100$, %	$K = 150$, %
HNF	54.05	58.23	**58.52**
HOGNMDH	56.49	**61.42**	**65.81**

Bold values indicate results of the proposed and best results

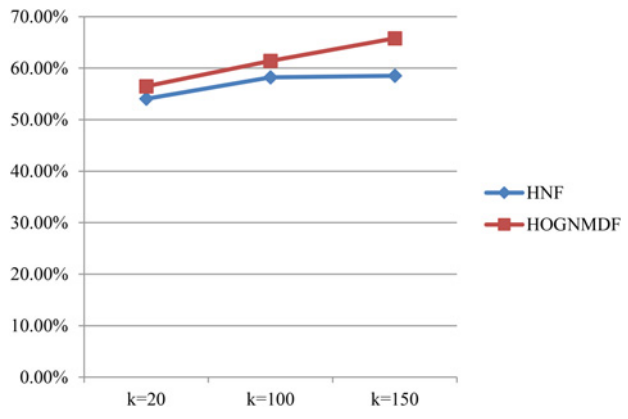

Fig. 5 *Comparison of HNF and HOGNMDH with different cluster numbers*

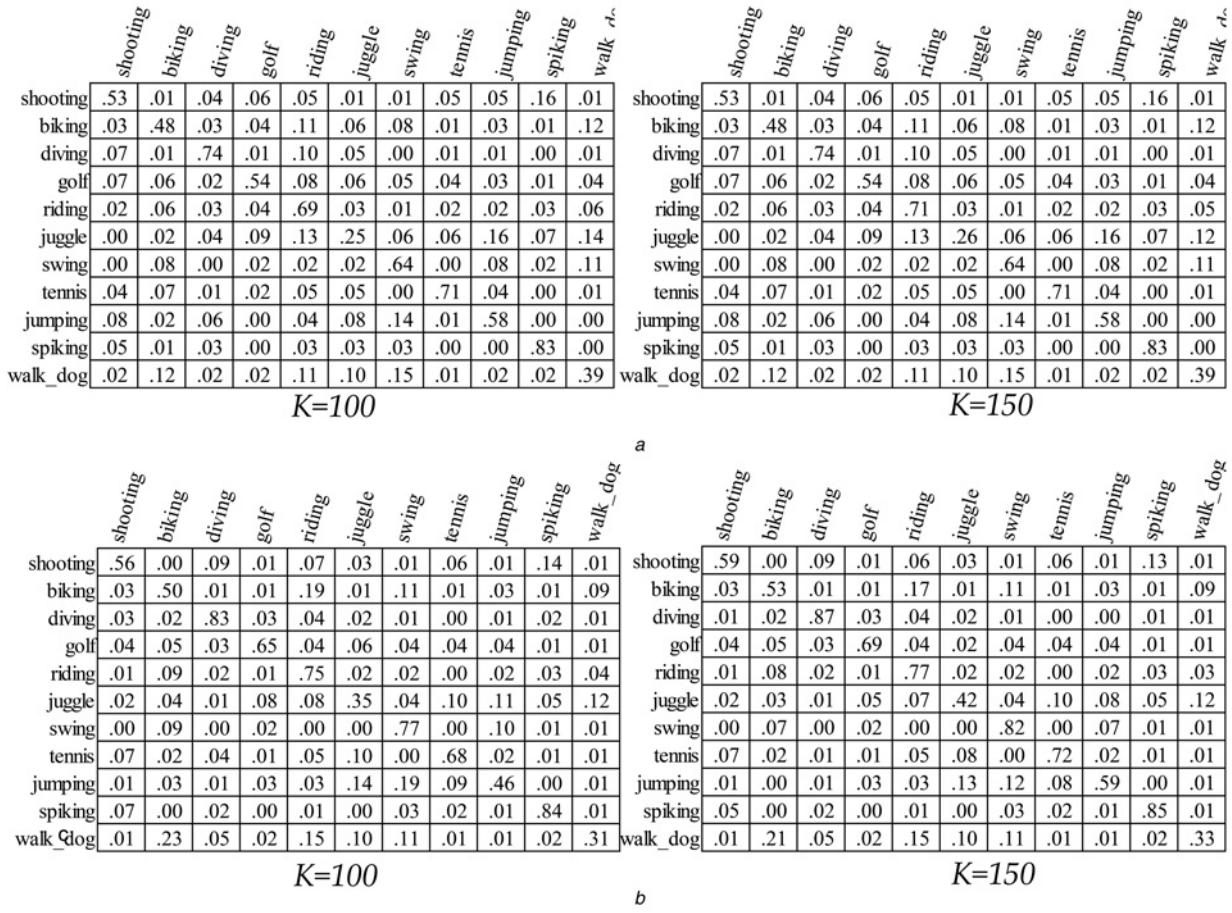

Fig. 6 *Confusion matrix of HNF and HOGNMDH*
a HNF feature
b HOGNMDH feature

From the comparison, we can find that compared with HOG and IMHcd, the improvement of MDH is more than 2%, and it is also better than MBH. Moreover, according the theory of feature descriptor in human motion recognition, the appearance feature combined with motion feature has better performance. In the experiment, we also compared the motion combine with appearance feature in Table 1.

In Table 1, HOG (histogram of orientation of gradient) is the appearance feature, HNF means the HOG feature combined with HOF. Moreover, HOGNMDF means the HOG feature combined with MDH. From the result, the performance of HOGNMDF is better than HNF, the improvement of HOGNMDF is more than 2%.

As mentioned in Section 3.2, the cluster number is sensitive to recognise performance. In the experiment, we discuss the cluster number k for action recognition. The value of k is set 20–150. Moreover, we have comprised the performance of HNF and HOGNMD feature. The experimental result is shown in Table 2 and Fig. 5.

From Table 2 and Fig. 5, we can find that the accuracy of HNF feature at $k = 100$ and $k = 150$ are 58.23 and 58.23%, respectively. The accuracy of HOGNMDH feature is 61.42 and 65.81%. The improvement of MDH at $k = 100$ and $k = 150$ is 3.19 and 7.89%, respectively. It verifies the effective of motion correction further. At the same time, there are almost have no improvement of HNF feature while the cluster number k increase from 100 to 150. Finally, the confusion matrix of the HNF feature and HOGNMDH is shown in Figs. 6*a* and *b*.

5 Conclusions

In this study, we propose a novel motion correction method and motion descriptor called MDH. In MDH, the camera motion is estimated, and relative motion is computed by the motion difference between the optical flow and the camera motion. To verify the effective of the proposed motion correction method, the MDH is built to recognise human motion. Experimental results by comparison with other relative motion descriptors show that the proposed descriptor is effective in motion description with camera movement. The motion correction method is useful to estimate real motion in camera movement scene. MDH is generally for other action recognition approaches and other vision tasks. In the future, we will use more a robust and discriminative action recognition approach to achieve better performance.

6 Acknowledgments

The work was supported by the Nature Science Foundation of China (no. 61502182), the Natural Science Foundation of Fujian Province of China (nos. 2014J01249, 2015J01253).

7 References

[1] Pavlovic V.I., Sharma R., Huang T.S.: 'Visual interpretation of hand gestures for human–computer interaction: a review', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, (7), pp. 677–695
[2] Phade G.M., Uddharwar P.D., Dhulekar P.A., *ET AL.*: 'Motion estimation for human–machine interaction'. IEEE Int. Symp.

Signal Processing and Information Technology, 2014, pp. 149–154
[3] Dawn D.D., Shaikh S.H.: 'A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector', *Vis. Comput.*, 2016, **32**, (3), pp. 1–18
[4] Aggarwal J.K., Ryoo M.S.: 'Human activity analysis: a review', *ACM Comput. Surv.*, 2011, **43**, (3), pp. 194–218
[5] Wang H., Schmid C.: 'Action recognition with improved trajectories'. IEEE Int. Conf. Computer Vision, 2013, pp. 3551–3558
[6] Bouguet J.Y.: 'Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm', *Acta Pathol. Jpn.*, 2000, **22**, (2), pp. 363–381
[7] Laptev I.: 'On space-time interest points', *Int. J. Comput. Vis.*, 2005, **64**, (2-3), pp. 107–123
[8] Xu L., Dai Z., Jia J.: 'Scale invariant optical flow'. European Conf. Computer Vision, Springer-Verlag, 2012, pp. 385–399
[9] Laptev I., Lindeberg T.: 'Local descriptors for spatio-temporal recognition', *Spat. Coherence Vis. Motion Anal.*, 2006, **3667**, pp. 91–103
[10] Wang H., Ullah, M.M., Klaser, A., *ET AL.*: 'Evaluation of local spatio-temporal features for action recognition'. 20th British Machine Vision Conf., BMVC, 2009
[11] Niebles J.C., Chen C.W., Li F.F.: 'Modeling temporal structure of decomposable motion segments for activity classification'. European Conf. Computer Vision, ECCV, 2010, pp. 392–405
[12] Liu J., Luo J., Shah M.: 'Recognizing realistic actions from videos in the Wild'. 2009 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops, CVPR Workshops, 2009