Learning a Discriminative Feature Descriptor with Sparse Coding for Action Recognition

Lingqiao Li[1,2], Tao Zhang[3], Xipeng Pan[1], Huihua Yang[1,2,*], Zhenbing Liu[2]
[1]School of Automation, Beijing University of Posts and Telecommunications, Beijing, China
[2]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China
[3]Department of Computer Science and Technology, Jiangnan University, Wuxi, China
Corresponding author: Huihua Yang (e-mail: yhh@bupt.edu.cn).

Abstract—**In this paper, we propose a novel algorithm for action recognition. The contribution of our work is three-fold. First, modified Weber local descriptor (IWLD) is proposed to capture the form cues of the action video sequences. Through introducing novel Weber magnitude and orientation components, our proposed IWLD can represent local patterns more effectively and accurately than existing Weber local descriptor (WLD). Second, to describe the form feature, histogram of improved Weber orientation Magnitude (HIOWM) is constructed. Considering motion and context cues also have discriminative power, we further propose a scheme that fuses HIOWM with motion and context cues to generate motion context HIOWM (MCHIOWM) descriptor to represent action video sequences. Third, for the sake of the more discriminative feature, we adopt sparse coding method to further refine the selected MCHIOWM. We present experiments to validate that the proposed framework obtains the competitive performance compared with the state-of-the-art methods.**

Keywords-action recognition; sparse coding; weber descriptor

I. INTRODUCTION

As one of the hot research areas in machine learning, action recognition based on computer vision has been generally used in some aspects. Because of the large changes in action types, such as different posture and body size in video data, action recognition remains a challenging problem. While salient low-level features have become one of the key issues for reports in both image processing [1] and pattern recognition [2]-[3]. For instance recently, Weber's law reveal a fact: human perception of one target depends on not only the variation of a stimulus but also the original intensity of the stimulus, the Weber Local Descriptor (WLD) can be used to represent the characteristics of the local area [4] and has been applied in object recognition region [5].

To achieve a robust detection and recognition for human behavior, two issues include robust actions representation and actions classification should be addressed. Many representation methods have recently been proposed, the most representative methods can be divided into two categories: holistic feature and local feature. Local feature representation methods mainly rely on the spatiotemporal interest points [6]-[7]. In [8], Savarse et al. designed spatiotemporal correlogram, which can make flexible long-range temporal information change into the spatiotemporal motion pattern. Hierarchically considering spatiotemporal relationship of feature, Ryoo et al. [9] proposed to use a novel matching algorithm to measure the similarity between different local features.

In addition, In [10], Ding et al. presented a hierarchical method for action representation more accurately. At low level, for the interest point, a novel feature descriptor is constructed to capture spatial information about continuous motion segments. And at high level, to incorporate spatial and temporal information, a kind of continuous motion feature descriptor is depicted.

In our paper, we propose a novel feature extraction descriptor for human action representation and a new sparse model for performing action classification effectively by using this novel holistic descriptor. The remainder of this paper is as follows. Section 2 introduces our proposed algorithm. Section 3 demonstrates our experimental results. Section 4give the conclusion of this paper.

II. OUR PROPOSED METHODS

We first describe improved Weber local descriptor and then by utilizing this improved WLD; histogram of improved Weber orientation (HIWO) is computed. Secondly, based on HIWO, the MCHIWO descriptor combining HIWO with motion and context cues is presented for action representation. At last, we adapt sparse coding to reduce feature dimension of low-level descriptors.

A. Modified Weber local descriptor

Weber Local Descriptor (WLD) describes one phenomenon: only if the ratio of the change of a stimulus to original stimulus is large enough, any change can be watched. Original descriptor was described as [4]:

Weber Magnitude:

$$\xi_m(x_c) = \arctan(\alpha \sum_{i=0}^{p-1} \frac{x_i - x_c}{x_c}) \tag{1}$$

where xc denotes the center pixel of xi, where each element was sampled from x_0 to x_{p-1} (p is the neighborhood size). α is a parameter adjustment factor. If $\xi_m(x_c)$ is near zero, it usually denotes a flat area [5].

Weber Orientation:

$$\xi_o(x_c) = \arctan(\frac{x_1 - x_5}{x_3 - x_7}) \tag{2}$$

where x_1-x_5 and x_3-x_7 represents the intensity difference respectively.

However, Weber magnitude in Eq.(1) do not consider changing orientations for eight $(x_i$-$x_c)/x_c$, and simply

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00030

regard these as values. So it can not reflect the intensity change accurately. Particularly, when a point is in non-flat area, Eq. (1) may give $\xi_m(x_c)$ very small value, which corresponds to a flat area [5].

Therefore, we propose to modify Weber local descriptor (IWLD):

Weber Difference in *x*:

$$\xi_{m-x}(x_c) = \arctan \alpha \sum_{i=0}^{p-1} \frac{x_i - x_c}{x_c} \cos \vartheta_i \tag{3}$$

Weber Difference in *y*:

$$\xi_{m-y}(x_c) = \arctan \alpha \sum_{i=0}^{p-1} \frac{x_i - x_c}{x_c} \sin \vartheta_i \tag{4}$$

Novel Weber Magnitude:

$$\xi_m(x_c) = \sqrt{\xi_{m-x}(x_c)^2 + \xi_{m-y}(x_c)^2} \tag{5}$$

Novel Weber Orientation:

$$\xi_o(x_c) = \arctan(\frac{\xi_{m-y}(x_c)}{\xi_{m-x}(x_c)}) \tag{6}$$

Where ϑ_i denotes the angle between *x* and (x_i-x_c)/x_c direction. By using $\xi_{m-x}(x_c)$ and $\xi_{m-y}(x_c)$, proposed Weber magnitude and orientation components can be got through (5) and (6). Compared to WLD, IWLD can depict local patterns more accurately. Eq. (5) gives a response value.

B. MCHIOWM Descriptor

Local descriptor. Input image windows were fixed as size *M*×*N*. No foreground segmentation (silhouette) is required. Each frame descriptor is a concatenation of a histogram of the improved Oriented Weber Magnitude (HIOWM) and a histogram of the optic flow (HOF) inside the image window.

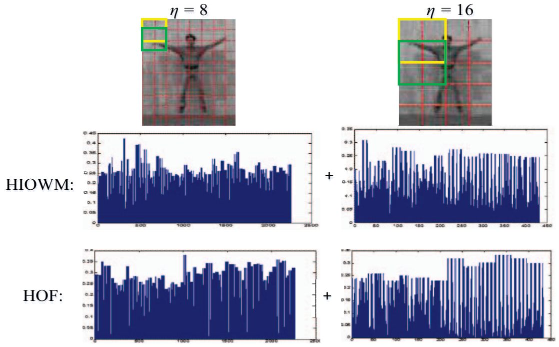


Figure 1. Constructing HIOWM and HOF descriptor.

HOG (histograms of oriented gradients) and HOF [11] [12] have achieved excellent results for many action recognition tasks; we propose to construct histograms for IWLD map of each frame which is similar to method described by [11]. When constructing HIOWM descriptor, for each block, *ς, η, β* are parameters. The function of the novel Weber magnitude is introduced to our algorithm. Each block is normalized independently with their *L2* norm. The blocks are typically overlapped as 0.5. Taking the spatial property of the local shape into account, we set *η* = 8 and *η* = 16 as varying resolution levels in the implementation. The final descriptor HIOWM for the image is a concatenation of two HIOWM vectors at *η* = 8 and *η* = 16, as illustrated in Fig. 1. Other default parameters for our experiments are *ς* = 2 and *β* = 9, which showed to give best performance when cross validating on the training set of Weizmann.

We employed the Lucas-Kanade algorithm [13] to get dense optical flow. When constructing HOF descriptor, we set parameters for HOF similar to that for HIOWM except to set last bin as zero bins [11], these pixels whose optical flow magnitudes are lower than a threshold are saved. By concatenating HIOWM and HOF, we obtain a novel frame descriptor Motion HIOWM (MHIOWM).

Considering context is important cue for action recognition, we combine each current frame descriptor with previous ones as the context information. For each frame of them, we construct MHIOWM descriptor and then stack them together into a context descriptor. For each current frame, its MHIOWM descriptor is also converted into the first q_f principal components by the same method. The extracted *N*-dimensional context descriptor is then added to *N*-dimensional frame descriptor, which constitute Motion Context HIOWM (MCHIOWM). In our experiment, we set the parameter *L* = 5, q_c = 250 and q_f = 500. As is detailed in Fig. 2.

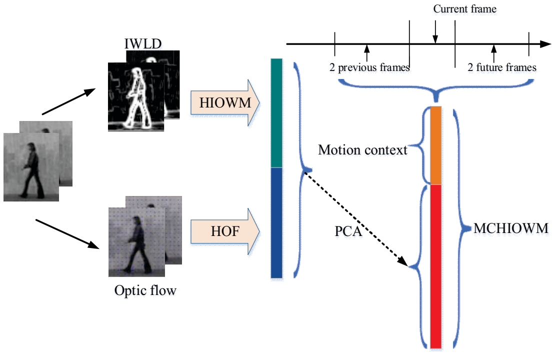


Figure 2. Constructing MCHIOWM.

C. KDE and Sparse coding scheme-based feature selection

MCHIOWM descriptor is a high-dimensional feature containing some redundant features. To keep the efficiency of calculation, the KDE-based approach [14][15] is used to select the most discrimitive features.

In order to get a discrimitive feature, the common Gaussian kernel density estimator [15] can be used. However, it lacks local adaptivity, so in order to reduce ambiguity and increase adaptability, K(•) is chosen to be an adaptive kernel, as described in [14].

Let *X* denotes MCHIOWM feature vectors, *Y* = *[y_1 ,y_2 ,...,y_N](Y ∈ R)*, where y_i denotes *i*-th vector. Thus, we can get the following sparse coding problem:

$$Z = \arg\min_{Z \in R} \frac{1}{2} \|Y - DZ\|_{\ell_2}^2 + \lambda \|Z\|_{\ell_1} \tag{7}$$

Where *Z = [Z_1, Z_2,...,Z_N](Z ∈R)* and Z_i denotes the sparse representation of vector y_i . *D = [d_1, d_2,...,d_N](D ∈R)* belong to a pre-trained dictionary. $\lambda$is a positive regularization parameter (it is set to 0.069 in [16]). In general, we can use the LARS-lasso approach [17] to get a sparse *Z* in Eq. 7.

Let *X = [x_1, x_2,..., x_N](X ∈R)* denote the reduced MCHIOWM features, then we can define the dictionary learning problem in the following manner:

$$\mathrm{Z} = \arg\min_{U \in R} \frac{1}{M} \sum \frac{1}{2} \left\| x_i - Du_i \right\|_{\ell_2}^2 + \lambda \left\| u_i \right\|_{\ell_1} \quad (8)$$

Where *C* is a convex set and *U = [u1, u2,...,uN](U ∈ R)* , so we can get:

$$\mathrm{C} \quad \left\{ D \in \mathbb{R}, s.t. \left\| d_i \right\|_{\ell_i} \leq 1, i \in \{1, \ldots, k\} \right\} \quad (9)$$

However, Eq. 9 is not convex when *D* and *U* are not constant value. We use online dictionary learning algorithm [16] to solve this problem.

To further optimize the relatively redundant feature, we use the max pooling method, which outperforms the average pooling [18][19]. In classification stage, the SVM with RBF kernel approach is used.

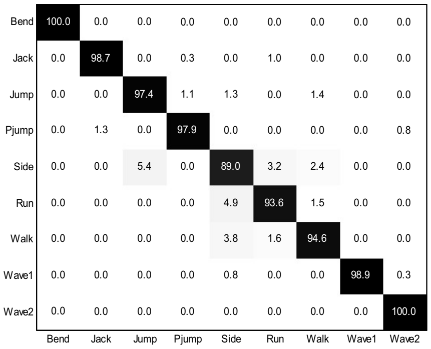
III. Results and Discussion

To prove the effectiveness of our proposed approach, we conducted experiment on two public dataset: Weizmann dataset [20] and KTH [20] dataset.

A. Experiments on the Weizmann dataset

We evaluate our algorithm on Weizmann dataset by the leave-one-out cross-validation method: 8 subjects are used to train, the others are used to test; each experiment is performed nine times and the final results are averaged.

The corresponding experimental results can be shown in Fig. 3. As illustrated in the figure, our algorithm obtains the best performance.



	Bend	Jack	Jump	Pjump	Side	Run	Walk	Wave1	Wave2
Bend	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Jack	0.0	98.7	0.0	0.3	0.0	1.0	0.0	0.0	0.0
Jump	0.0	0.0	97.4	1.1	1.3	0.0	1.4	0.0	0.0
Pjump	0.0	1.3	0.0	97.9	0.0	0.0	0.0	0.0	0.8
Side	0.0	0.0	5.4	0.0	89.0	3.2	2.4	0.0	0.0
Run	0.0	0.0	0.0	0.0	4.9	93.6	1.5	0.0	0.0
Walk	0.0	0.0	0.0	0.0	3.8	1.6	94.6	0.0	0.0
Wave1	0.0	0.0	0.0	0.0	0.8	0.0	0.0	98.9	0.3
Wave2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

Figure 3. Confusion matrix on the Weizmann dataset.

Fig.4 illustrates that the proposed context descriptor improve the performance of our framework. It can be seen that if a single frame descriptor (MHIOWM) is adopted, the recognition rate is only 93.6%. However, the recognition rate reduced to 96.7% when we use 5 frames around the current frame to generate context descriptor and append this context descriptor to current frame descriptor so as to form Motion Context HIOWM (MCHIOWM). It demonstrates that context information provides important discriminative cue for action recognition.

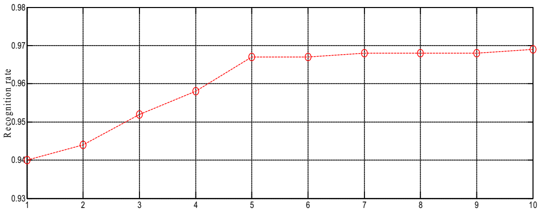


Figure 4. The influence of w in the classification scheme.

To make the comparison fair, all the features are combined with our proposed modified sparse framework for action classification. From the Table I, we can see that every action feature owns discriminative power for action recognition. The MHOWM descriptor, whose form feature is constructed on the WLD map, outperforms the traditional HOG+HOF descriptor. It turns out that the MCHIOWM feature, that incorporates the motion context information, achieves the highest recognition rate.

TABLE I. Contribution of Proposed Features

Action feature	Accuracy (%)
HOG+HOF	90.2
HOWM+HOF(MHOWM)	92.0
HIOWM+HOF (MHIOWM)	93.6
CONTEXT+HIOWM+HOF (MCHIOWM)	96.9

To demonstrate the robustness of our proposed algorithm, we make comparisons with other classifier. The performance is robust, as can be seen in Fig. 5. Experiments show that on the Weizmann dataset our method gains the state-of-the-art result as illustrated in Table II.

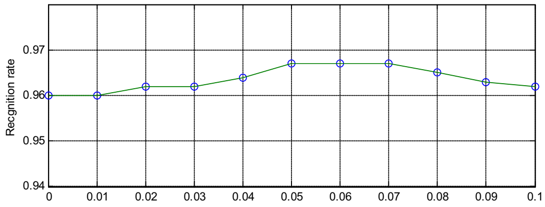


Figure 5. The influence of w in the classification scheme.

TABLE II. Comparison with Previous Work on the Weizmann Dataset

Method	Accuracy (%)
Our method	96.9
Bregonzio et al.[12]	96.7
Junejo et al. [21]	95.33
Ding et al. [10]	96.7
G. Lu et al. [22]	95.6
Chen et al. [23]	95.7
Hou et al. [1]	96.6

B. Experiments on the KTH dataset

To further verify our algorithm, we also evaluated it on the KTH dataset. Fig. 6 shows the average confusion matrix. In this figure, we can see that four action classes (out of six in totals) are perfectly detected. The most difficulty two action classes are "Jog" and "Run". This confusion happens reasonably, since they are quite similar. However, we cannot distinguish action classes “wave” and “clap”, which seems the same.

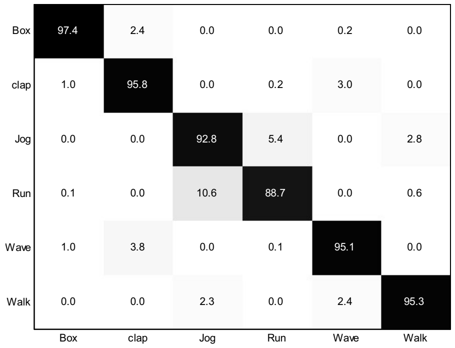


Figure 6. Confusion matrix on the KTH dataset.

Table III illustrates the comparison results on HOG+HOF, HWOM+HOF, and MCHIOWM descriptors. Each of them offers powerful discriminability. MCHIOWM descriptors provide a better feature representation. Also in Weizmann dataset, our proposed algorithm achieves higher recognition rate than SVM or SRC. Experiments show that our sparse representation model obtains the competitive results, as summarized in Table IV.

TABLE III. CONTRIBUTION OF PROPOSED FEATURES

Action feature	Accuracy (%)
HOG+HOF	88.4
HWOM+HOF	89.2
MCHIOWM	94.2

TABLE IV. COMPARISON WITH PREVIOUS WORK ON THE KTH DATASET

Method	Accuracy (%)
Our method	95.20
Bregonzio et al[12]	91.80
Junejo et al. [21]	93.60
Ding et al. [10]	93.3
G. Lu et al. [22]	93.90
Chenet al. [23]	93.36
Hou et al. [1]	95.09

IV. CONCLUSIONS

This paper introduced a novel action recognition approach based on the MCHIOWM feature and sparse coding. Compared with state-of-the-arts, the proposed method performs the best.

ACKNOWLEDGMENT

This work was supported in part by the NSFC under Grant 61702226, 61562013 and 61672263, the National Key Research and Development Program of Guangxi Province under Grant AB16380293, the Natural Science Foundation of Guangxi Province under Grant 2017GXNSFDA198025.

REFERENCES

[1] Y. Hou, Z. Li, P. Wang, et al, “Skeleton Optical Spectra Based Action Recognition Using Convolutional Neural Networks,” IEEE Trans. Circuits and Systems for Video Technology, vol. 28, pp. 1-8, 2018.

[2] W. Bian, D. Tao, and Y. Rui, “Cross-domain human action recognition,” IEEE Trans. Systems Man and Cybernetics Part B Cybernetics, vol. 42, pp. 298–307, 2012.

[3] K. Fahad Shahbaz, W. Joost Van De, R. MA, et al, “Scale coding bag of deep features for human attribute and action recognition,” Machine Vision and Applications, vol. 29, pp. 55-71, 2018.

[4] J. Chen, S. Shan, C. He, et al, “WLD: A robust local image descriptor,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, pp. 1705–1720, 2010.

[5] B. Wang, W.F. Li, W.M. Yang, et al, “Illumination normalization based on Weber’s law with application to face recognition,” IEEE Signal Processing Letters, vol. 18, no. 8, pp. 462–465, 2011.

[6] D. Debapratim Das and S. H. Shaikh, “A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector,” Visual Computer, vol. 32, no. 3, pp. 289-306, 2016.

[7] Y. Li, R. Xia, Q. Huang, et al, “Survey of Spatio-Temporal Interest Point Detection Algorithms in Video,” IEEE Access, vol. 5, no. 99, pp. 10323-10331, 2017.

[8] S. Savarese, A. DelPozo, and J. Niebles, “Spatial-temporal correlatons for unsupervised action classification,” Proc. IEEE WMVC, pp. 1–8, 2008.

[9] M. Ryoo and J. Aggarwal, “Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities,” Proc. IEEE 12th Int’l Conf. Computer Vision, pp. 1593–1600, 2009.

[10] W. Ding, K. Liu, F. Cheng, et al, “Learning hierarchical spatio-temporal pattern for human activity prediction,” J.Visual Communication and Image Representation, vol. 35, pp. 103-111, 2016.

[11] I. Laptev, M. Marszalek, C. Schmid, et al, “Learning realistic human actions from movies,” Proc. IEEE Int’l Conf. CVPR, pp. 1–8, 2008.

[12] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” Proc. ICCV, pp. 886–893, 2005.

[13] BD. Lucas, and T. Kanade, “An iterative image registration technique with an application to stero vision,” IJCAI, vol. 2, no. 3, pp. 121–130, 1981.

[14] X. Geng, and G. Hu, “Unsupervised feature selection by kernel density estimation in wavelet-based spike sorting,” Biomedical Signal Processing & Control, vol. 7, no. 2, pp. 112–117, 2012.

[15] ZI. Botev, JF. Grotowski, and DP. Kroese, “Kernel density estimation via diffusion,” The Annals of Statistics, pp. 2916–2957, 2010.

[16] J. Mairal, F. Bach, J. Ponce, et al, ″Online dictionary learning for sparse coding,” Annual International Conference on Machine Learning, pp. 689–696, 2009.

[17] J. Yang, K. Yu, and T. Huang, “Supervised translation-invariant sparse coding,” IEEE Computer Vision & Pattern Recognition, vol. 26, no. 2, pp. 3517–3524, 2010.

[18] FDMD. Souza, G. C. Chavez, EADV. Jr, et al, “Violence detection in video using spatio-temporal features,” Graphics Patterns & Images, pp. 224–230, Sept. 2011.

[19] Y. L. Boureau, J. Ponce, and Y. Lecun, “A theoretical analysis of feature pooling in visual recognition,” International Conference. Machine Learning, vol. 32, no. 4, pp. 111–118, 2010.

[20] M. Blank, L. Gorelick, E. Shechtman, et al, “Action as space–time shapes,” IEEE. ICCV, 2005.

[21] I. Junejo, E. Dexter, I. Laptev, et al, “View-independent action recognition from temporal self-similarities,” IEEE Trans. Pattern Anal & Mach Intell1, vol. 33, no. 1, pp. 172–185, 2010.

[22] G. Lu, and M. Kudo, “Learning action patterns in difference images for efficient action recognition,” Computer and Robot Vision, pp. 328-336, 2014.

[23] C. Chen, K. Liu, and N. Kehtarnavaz, “Real-time human action recognition based on depth motion maps,” J.Real-Time Image Processing, pp. 156-163, Dec. 2016