LETTER

Statistics on Temporal Changes of Sparse Coding Coefficients in Spatial Pyramids for Human Action Recognition

Yang LI[†], ***Student Member***, **Junyong YE**[†a)], **Tongqing WANG**[†], ***and*** **Shijian HUANG**[†], ***Nonmembers***

SUMMARY Traditional sparse representation-based methods for human action recognition usually pool over the entire video to form the final feature representation, neglecting any spatio-temporal information of features. To employ spatio-temporal information, we present a novel histogram representation obtained by statistics on temporal changes of sparse coding coefficients frame by frame in the spatial pyramids constructed from videos. The histograms are further fed into a support vector machine with a spatial pyramid matching kernel for final action classification. We validate our method on two benchmarks, KTH and UCF Sports, and experiment results show the effectiveness of our method in human action recognition.

key words: *sparse coding, temporal changes, spatial pyramid, human action recognition*

1. Introduction

Recognition of human actions from videos has gained much interest in recent years due to its wide ranges of potential applications, such as video surveillance, video retrieving, and human-computer interaction. Although much progress has been made, it still remains a challenging problem due to cluttered background, camera motion, occlusions, viewpoint changes, and large variations in the same class.

Recently, it has become more and more popular to employ sparse representation-based methods for various computer vision tasks, such as image classification [1], face recognition [2], [3], and human action recognition [4]–[8]. Sparse representation-based methods for human action recognition first compute the sparse feature representation with learned dictionary and then pool over the entire video to form the final representation, where average pooling and max pooling are usually used. However, pooling over the entire video neglects any spatio-temporal information of features, resulting in non-discriminative representation.

A common way to overcome this drawback is to use spatio-temporal pyramids [9]. However, this provides a representation too coarse to capture the rich relationship between features. Moreover, it needs shot boundary detection which is not robust enough for complex video contents.

To robustly capture the spatio-temporal information of features, we propose a novel method based on statistics on changes of sparse coding coefficients. We first represent features using the sparse coding method. Then we count the number of changes of sparse coding coefficients frame by frame to robustly capture the temporal information of features in spatial pyramids. Finally, the statistics histograms are fed into a support vector machine with a spatial pyramid matching kernel [10] for final classification. Our method needs no additional steps, such as shot boundary detection, and therefore is more robust. Moreover, our method is easy to compute. We test our method on KTH dataset [11] and UCF Sports dataset [12], and experiment results show its effectiveness in human action recognition.

The remainder of the letter is organized as follows. In Sect. 2, we first present the sparse coding method and then introduce our proposed method in detail. In Sect. 3, we conduct experiments on two benchmark datasets to demonstrate the effectiveness of our method. Finally, in Sect. 4, we conclude the letter.

2. Proposed Approach

In this section, we first introduce the sparse coding method and then present how to count the changes of sparse coding coefficients. At last, we show how to use our method for human action recognition.

2.1 Sparse Coding

There are many methods can generate sparse representation for features of a video. Although our method is independent on sparse coding methods, considering the coding efficiency, we employ LLC [1] as our feature coding method.

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{P\times N}$ be a set of N features and $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K] \in \mathbb{R}^{P\times K}$ the learned dictionary with K atoms. Let $\mathbf{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N] \in \mathbb{R}^{K\times N}$ be the coding coefficient matrix. LLC coding method computes coding coefficients as follows:

$$\boldsymbol{\alpha}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|^2 + \lambda\|\mathbf{e}_i \odot \boldsymbol{\alpha}_i\|^2 \quad s.t. \quad \mathbf{1}^T\boldsymbol{\alpha}_i = 1 \tag{1}$$

where $\mathbf{e}_i = \exp(\mathrm{dist}(\mathbf{x}_i, \mathbf{D})/\sigma)$ and $\mathrm{dist}(\mathbf{x}_i, \mathbf{D})$ denotes the Euclidean distance between $\mathbf{x}_i$ and atoms of dictionary $\mathbf{D}$. σ is a parameter controlling the weight vector $\mathbf{e}_i$. $\odot$ denotes the element-wise multiplication. In our experiments, we employ the approximated LLC for fast encoding, i.e., each feature is encoded by the nearest k dictionary atoms.

Manuscript received February 6, 2015.
Manuscript revised May 6, 2015.
Manuscript publicized June 1, 2015.

[†]The authors are with Key Laboratory of Optoelectronic Technology and Systems of the Ministry of Education, Chongqing University, Chongqing, China.
a) E-mail: ygyocr@cqu.edu.cn

DOI: 10.1587/transinf.2015EDL8037

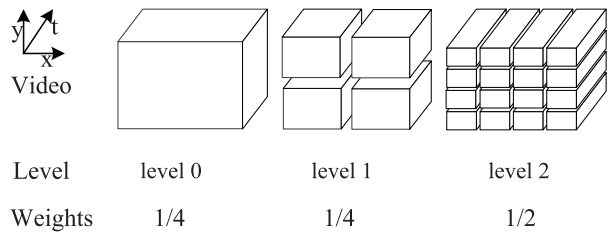

Fig. 1 Illustration of constructing a spatial pyramid from a video

2.2 Statistics on Temporal Changes of Coding Coefficients

Traditional methods pool over the entire video, neglecting any spatio-temporal information. For human action recognition, the temporal information of features is very important. We propose to employ the temporal changes of coding coefficients to capture the temporal information between features.

Given a video, we divide it into 3D grids only in spatial space at L different levels of resolution, and a spatial pyramid is constructed. As illustrated in Fig. 1, we take $L = 2$ for example. The l-th level will have 4^l cells. In a cell with V frames, the representation $\mathbf{f}_i = [f_i(1), \ldots, f_i(K)]^T$ for frame v_i is obtained by pooling over the sparse coding coefficients of features located in frame v_i. Then temporal information for frame v_i can be described by changes of coding coefficients between v_i and its next frame v_{i+1}. The changes can be captured by a matrix $\mathbf{M}_i$ of size 2-by-K:

$$\mathbf{M}_i = \begin{bmatrix} m_i(1,1) & \ldots & m_i(1,K) \\ m_i(2,1) & \ldots & m_i(2,K) \end{bmatrix} \tag{2}$$

Each row of $\mathbf{M}_i$ consists of 0 and 1. The non-zero elements of its first row denote the coding coefficients of v_i is larger than that of v_{i+1}, and the non-zero elements of second row denote the coding coefficients of v_i is smaller than that of v_{i+1}, i.e.,

$$m_i(1,j) = \begin{cases} 1, & if\ f_i(j) > f_{i+1}(j) \\ 0, & otherwise \end{cases} \tag{3}$$

and

$$m_i(2,j) = \begin{cases} 1, & if\ f_i(j) < f_{i+1}(j) \\ 0, & otherwise \end{cases} \tag{4}$$

Then matrix $\mathbf{M}_i$ is reshaped to a vector $\mathbf{b}_i$ as the histogram representation for frame v_i. The histogram representation $\mathbf{g}$ of the cell is computed as:

$$\mathbf{g} = \sum_i \mathbf{b}_i / V^* \tag{5}$$

where V^* is the number of frames taken into consideration[†]. The concatenated histogram of cells at level l forms the final representation for the corresponding grid.

[†]A frame is considered only if it and its next frame have non-zero and different coding coefficients.

2.3 Action Classification

After histogram representation of all grids is computed, we employ a support vector machine with a spatial pyramid matching kernel for action classification. Suppose we have C channel features, the spatial pyramid matching kernel between videos Y and Z is computed as:

$$\kappa(Y,Z) = \sum_{l=0}^{L} w^l \sum_{c=1}^{C} \kappa(\mathbf{h}_Y^{l,c}, \mathbf{h}_Z^{l,c}) \tag{6}$$

where w^l is the weight for the l-th level grid and computed as:

$$w^l = \begin{cases} 1/2^L, & if \quad l = 0 \\ 1/2^{L-l+1}, & if \quad 1 \le l \le L \end{cases} \tag{7}$$

$\mathbf{h}_Y^{l,c} = [h_Y^{l,c}(1), \ldots, h_Y^{l,c}(j), \ldots]^T$ and $\mathbf{h}_Z^{l,c} = [h_Z^{l,c}(1), \ldots, h_Z^{l,c}(j), \ldots]^T$ are the histogram representations of the c-th channel features in the l-th level grid for videos Y and Z respectively, and $\kappa(\mathbf{h}_Y^{l,c}, \mathbf{h}_Z^{l,c})$ is the histogram intersection kernel:

$$\kappa(\mathbf{h}_Y^{l,c}, \mathbf{h}_Z^{l,c}) = \sum_j \min(h_Y^{l,c}(j), h_Z^{l,c}(j)) \tag{8}$$

where j is the index of histogram components.

3. Experiments

In this section, to demonstrate the effectiveness of our method, we conduct experiments and compare it with other excellent methods on two benchmarks: KTH dataset and UCF Sports dataset.

3.1 Parameter Settings

For video representation, we use the code provided by the author of [13] with the default parameter settings to locate spatio-temporal interest points which are further described by histogram of gradient (HOG) and histogram of flow (HOF) [14]. The HOG/HOF features are further preprocessed by principal component analysis and whitening into features with dimensions of 100 (C = 1). We construct a three-level spatial pyramid (L=2), and max pooling is employed as the pooling function. The dictionary $\mathbf{D}$ is computed using k-means algorithm and its size is empirically set to 1000 for KTH and UCF Sports datasets.

3.2 Experiment Results on KTH Dataset

The KTH dataset is a standard and popular benchmark for human action recognition. It has six action classes in total (e.g., clapping, running, walking), each of which is performed in four different scenarios by 25 subjects, resulting in a total of 599 videos. We follow recent evaluations on KTH dataset using the LOOCV strategy for classification,

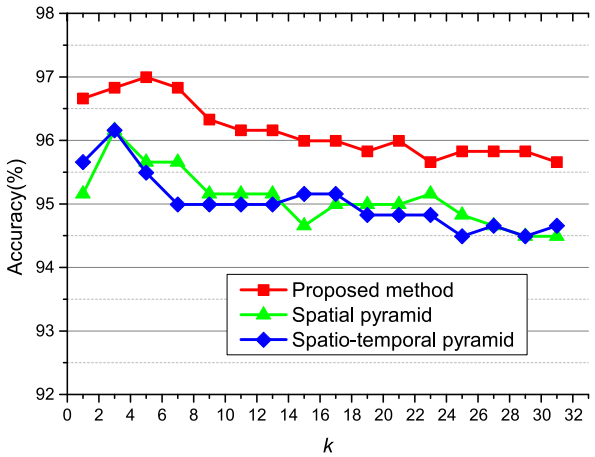

Fig. 2 Performance comparison with different k in LLC on KTH dataset

Table 1 Recognition results of different methods on KTH dataset

Method	Year	Accuracy(%)
Wang et al. [15]	2009	92.1
Zhang et al. [5]	2012	95.06
Wang et al. [8]	2012	94.17
Wang et al. [16]	2013	94.2
Peng et al. [7]	2014	94.4
Spatial pyramid + LLC		96.16
Spatio-temporal pyramid + LLC		96.16
Proposed method + LLC		96.99

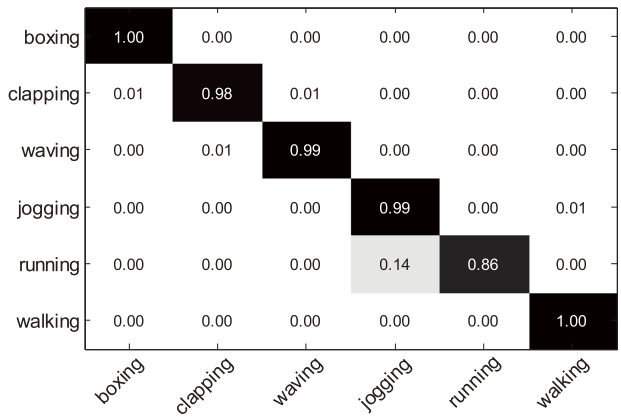

Fig. 3 Confusion matrix for KTH dataset

i.e., videos of twenty-four subjects are used as the training sets and the remainder as the testing sets.

In order to demonstrate the superiority, we compare the performance of our method with that of spatial pyramid method and spatio-temporal pyramid method. As illustrated in Fig. 2, with different k in LLC, our method always performs better. When $k = 5$, we achieve the best accuracy of 96.99%. Table 1 compares our result with previous works, and we achieve the highest accuracy using the same classification strategy. Figure 3 shows the confusion matrix for KTH dataset with $k = 5$. From it, we can observe that wrong classifications usually occur between "jogging" and "running", which are very similar and hard to classify.

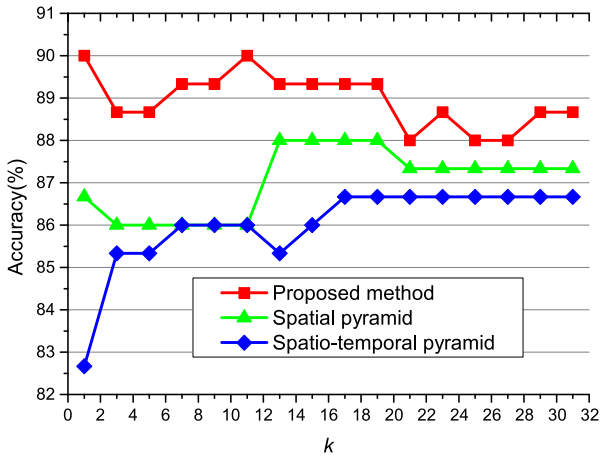

Fig. 4 Performance comparison with different k in LLC on UCF Sports dataset

Table 2 Recognition results of different methods on UCF Sports dataset

Method	Year	Accuracy(%)
Wang et al. [15]	2009	85.6
Zhang et al. [5]	2012	87.33
Wang et al. [8]	2012	86.6
Wang et al. [16]	2013	88.0
Zhang et al. [6]	2014	86.7
Spatial pyramid + LLC		88.0
Spatio-temporal pyramid + LLC		86.67
Proposed method + LLC		90.0

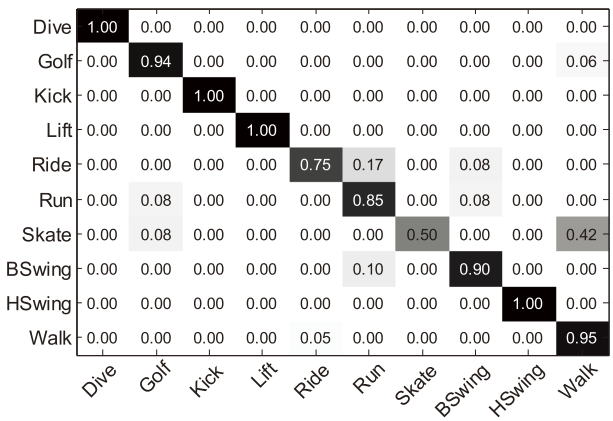

Fig. 5 Confusion matrix for UCF Sports dataset

3.3 Experiment Results on UCF Sports Dataset

The UCF Sports dataset consists of 150 video clips belonging to 10 action classes: diving (Dive), golf swinging (Golf), kicking (Kick), lifting (Lift), horse-riding (Ride), running (Run), skateboarding (Skate), swinging at the bench (BSwing), swinging at the high bar (HSwing), and walking (Walk). We increase the amount of samples and use a leave-one-sample out cross-validation setting as suggested in [15]. Figure 4 compares the performance of spatial pyramid method, spatio-temporal method, and our method with

different k in LLC. When $k = 1$ or $k = 11$, we achieve the best accuracy of 90.0%. Table 2 compares our result with other excellent methods, and we achieve the highest accuracy using the same classification strategy. Figure 5 shows the confusion matrix for UCF Sports dataset with $k = 11$. From it, we can observe that wrong classifications usually occur between "skateboarding" and "walking".

4. Conclusion

In this letter, we propose a novel histogram representation to overcome the drawback of traditional sparse representation-based methods for human action recognition. Our method counts the temporal changes of sparse coding coefficients of features in the spatial pyramids to incorporate spatio-temporal information of features. It needs no additional steps and is easy to compute. Experiment results on two benchmarks show the robustness of our method and its superiority in human action recognition.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities of China under Grant 10611201312014 and Scientific and Technological Research Program of Chongqing Municipal Education Commission of China under Grant KJ1401207.

References

[1] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality–constrained linear coding for image classification," Proc. 2010 IEEE Int. Conf. Comput. Vis. and Pattern Recognit., San Francisco, CA, pp.3360–3367, June 2010.

[2] M. Yang, D. Zhang, X.C. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," Proc. 2011 IEEE Int. Conf. Comput. Vis., Barcelona, Spain, pp.543–550, Nov. 2011.

[3] Q. Zhang, and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," Proc. 2010 IEEE Int. Conf. Comput. Vis. and Pattern Recognit., San Francisco, CA, pp.2691–2698, June 2010.

[4] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," Proc. 2010 10th Asian Conf. Comput. Vis., Queenstown, New zealand, pp.660–671, Nov. 2010.

[5] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Action recognition using context-constrained linear coding," IEEE Signal Process. Lett., vol.19, no.7, pp.439–442, 2012.

[6] S. Zhang, H. Yao, X. Sun, K. Wang, J. Zhang, X. Lu, and Y. Zhang, "Action recognition based on overcomplete independent components analysis," Inf. Sci., vol.281, pp.635–647, 2014.

[7] X. Peng, L. Wang, Y. Qiao, and Q. Peng, "A joint evaluation of dictionary learning and feature encoding for action recognition," Proc. 2014 22nd Int. Conf. Pattern Recognit., Stockholm, pp.2607–2612, May 2014.

[8] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," Pattern Recognit., vol.45, no.11, pp.3902–3911, 2012.

[9] J. Choi, W.J. Jeon, and S.-C. Lee, "Spatio-temporal pyramid matching for sports videos," Proc. 1st ACM Int. Conf. Multimed. Inf. Retr., Vancouver, BC, Canada, pp.291–297, Aug. 2008.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," Proc. 2006 IEEE Int. Conf. Comput. Vis. and Pattern Recognit., New York, United states, pp.2169–2178, June 2006.

[11] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," Proc. 17th Int. Conf. Pattern Recognit., Cambridge, England, pp.32–36, Aug. 2004.

[12] M.D. Rodriguez, J. Ahmed, and M. Shah, "Action mach: a spatiotemporal maximum average correlation height filter for action recognition," Proc. 2008 IEEE Int. Conf. Comput. Vis. and Pattern Recognit., Anchorage, AK, pp.1–8, June 2008.

[13] I. Laptev, "On space-time interest points," Int. J. Comput. Vis., vol.64, no.2-3, pp.107–123, Sept. 2005.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," Proc. 2008 IEEE Int. Conf. Comput. Vis. and Pattern Recognit., Anchorage, AK, pp.1–8, June 2008.

[15] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," Proc. 2009 20th Br. Mach. Vis. Conf., London, UK, pp.124.1–124.11, Sept. 2009.

[16] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," Int. J. Comput. Vis., vol.103, no.1, pp.60–79, 2013.