SPATIO-TEMPORAL PROXIMITY DISTRIBUTION KERNELS FOR ACTION RECOGNITION

Chunfeng Yuan[1], Weiming Hu[1], Hanzi Wang[2], Xi Li[1], Nianhua Xie[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] School of Computer Science, University of Adelaide, SA 5005, Australia

ABSTRACT

Spatio-temporal local feature based bag of visual words algorithm (BOVW) has shown promising results in complex human action classification. However, one key disadvantage of BOVW is geometrical unconstraint, which makes it impossible to recognize different actions with the same features but different spatial-temporal distribution of these features. In this paper, we exploit the spatio-temporal proximity distribution of local features in 3D space to characterize geometric context of action class. The obtained spatio-temporal proximity matrix models both the appearance and geometrics of human actions. Moreover, a spatio-temporal proximity distribution kernel (ST-PDK) is proposed to measure the similarity of videos, which satisfies Mercer condition and is directly incorporated into the kernel function of the SVM classifier. Our algorithm achieves the highest classification accuracy on KTH dataset, one of the most challenging and popular action datasets.

***Index Terms*—** action recognition, BOVW, spatio-temporal proximity distribution kernel

1. INTRODUCTION

Recently, local spatio-temporal features are widely used in action recognition tasks, because they are more robust to noise, occlusion, and geometric variations than global (or large-scale) features. Most of the previous works [1-4, 6] treat these spatial temporal features as a bag of visual words (BOVW), which is geometrical unconstrained omitting any long range or global information in either spatial or temporal domain. However, from Fig.1, it can be seen that the geometrical distribution of local features regularly varies among different action classes, and thus spatio-temporal information is very helpful for improving the action recognition accuracy. Therefore, our goal is to exploit the complex geometric relationships among local features in order to improve the recognition performance.

In object recognition of image, co-occurrence matrix [8, 9] and Proximity Distributions [10] of visual words are proposed to use as a global image descriptor for capturing the geometric information. In [10], the minimum of proximity matrices are used as the proximity distribution kernel to measure the distances of images. In [8] the stationary distribution derived from the normalized co-occurrence matrix forms the so-called Markov stationary features (MSF), and then the $\chi 2$ distance of MSF is adopted as the kernel of SVM classifier. In [9], the four co-occurrence matrices are calculated for four selected direction and hence Directed Markov Stationary Features (DMSF) are introduced.

In [8-10], it is demonstrated that both co-occurrence matrix and Proximity matrix are able to improve recognition performance in images. Inspired by these, we construct spatio-temporal proximity matrices in 3D space to represent the geometric information of videos. Besides, we compare the spatio-temporal proximity matrix with the spatio-temporal co-occurrence matrix deduced from spatial co-occurrence matrix [8, 9], theoretically and experimentally. Further, a spatial-temporal proximity distribution kernel (ST-PDK) is designed to use as SVM kernel. Experiments demonstrated that ST-PDK can effectively enforce and exploit geometric consistency among actions in the same category.

The remainder of the paper is organized as follows. Section 2 introduces how to produce ST-PDK in details. Section 3 illustrates the SVM classification approach based on ST-PDK. Section 4 reports experimental results on KTH human action datasets. Section 5 concludes the paper.

2. SPATIO-TEMPORAL PROXIMITY DISTRIBUTION OF LOCAL FEATURES

First of all, a large set of local features are used to represent each video sequence. We employ the Dollár et al.'s detector [7] to detect cuboids at every frame for each video and use the covariance descriptor [14] to describe the detected cuboids. Further, the features from training videos are quantized to form an appearance codebook (i.e. BOVW) by using the k-mean clustering method. Let $\{v_1,\ldots,v_K\}$ be K visual words of codebook. Fig.1 demonstrates the localization of the detected local features for six action video sequences in KTH dataset. It shows that the spatio-temporal information will be a very useful cue to improve the action recognition accuracy. To overcome the geometrical unconstraint of the ordinary BOVW approach, we propose a spatio-temporal proximity matrix as a global

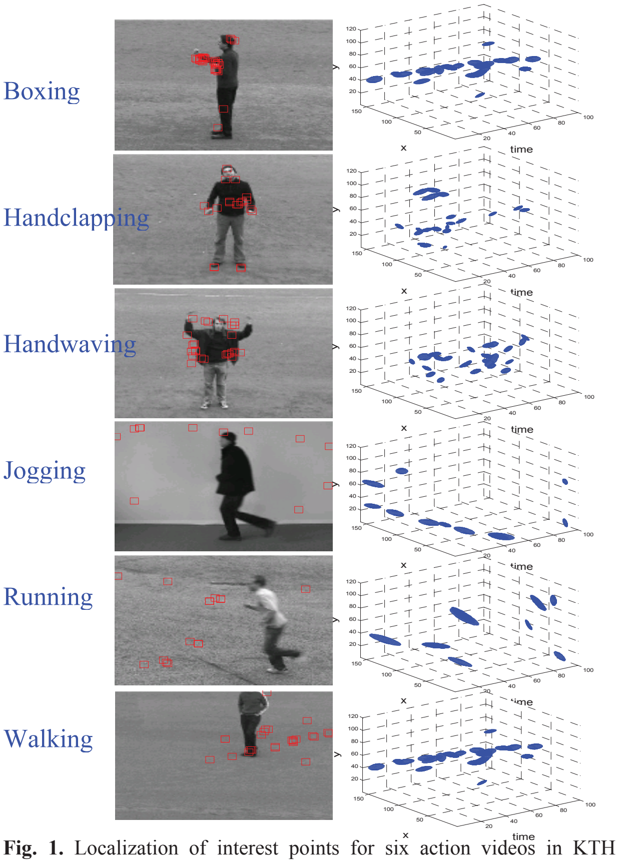


Fig. 1. Localization of interest points for six action videos in KTH dataset. In the left column, one key frame for each video is shown and all interest points detected in that video are overlapped on the key frame, which represents the spatial distribution of features. The right column represents the distribution of the features in spatio-temporal coordinates.

descriptor to characterize the complex geometric relationships among local features.

2.1. Spatio-temporal Proximity Matrix of Local Features

The Proximity Distributions [10] of local features was recently proposed to exploit the spatial relationships of local features, which shows that it has potential to improve the performance of object classification in images. Therefore we investigate the Proximity Distributions in video and construct proximity matrix in spatial-temporal domain.

Each video V is denoted as $\{(x_i, \alpha_i)\}_{1 \le i \le M}$, where x_i is the spatio-temporal position vector of the i^{th} detected local feature and α_i is the index of visual words. M is the total number of local features detected in the video. Given a video, the proximity matrix is defined as $H^r=(h_{i\,j})\in R^{K\times K}$. The element $h_{i\,j}$ is the number of the features belonging to the type v_i that are within the r-nearest neighbors of a feature in type $v_{j.}$

$$H^{r}(i,j)=h_{ij}=\#\{(\alpha_l,\alpha_m)\big|\alpha_l=i,\alpha_m=j,d_{NN}(x_l,x_m)\le r\} \quad (1)$$

where $d_{NN}(x_l,x_m)\le r$ indicates that x_m is within the r^{th} nearest neighbors of x_l, and the # means the number of feature pairs satisfying all the conditions listed in the brackets in Eq(1).

Moreover, the number of local feature pairs is instable in each video, because it relies heavily on video duraion and resolution. Hence, we normalize spatio-temporal proximity matrix by the addition of its elements:

$$H(i,j)=H^{r}(i,j)\Big/\textstyle\sum_{i=1}^{K}\sum_{j=1}^{K}H^{r}(i,j) \quad (2)$$

where $H(i,j)$ is the element of the normalized spatio-temporal proximity matrix, and $H=(H(i,j))_{K\times K}$ is used as the final spatio-temporal proximity matrix.

2.2. Comparison of Spatio-temporal Proximity and Co-occurrence Matrix

According to [8, 9], the spatio-temporal co-occurrence matrix can be obtained by extending the spatial co-occurrence matrix to the spatial-temporal domain. Spatio-temporal co-occurrence matrix is defined as $C = (c_{i\ j})\in R^{K\times K}$ with each element as

$$C(i,j)=c_{ij}=\#\{(\alpha_l,\alpha_m)\big|\alpha_l=i,\alpha_m=j,\left\|x_l-x_m\right\|\le d\} \quad (3)$$

where α_l and α_m are a pair of neighboring features with the distance not larger than d. By the same way, the spatio-temporal co-occurrence matrix is normalized too.

Obviously, the difference between spatio-temporal proximity and co-occurrence matrix is only the definition of the neighboring relationship between feature pairs. Spatio-temporal proximity matrix captures rank information between visual words, which is more reliable and sparse. In comparison, spatio-temporal co-occurrence matrix uses the difference of the absolute positions to measure the distribution of the distances between all pairs of visual labels. The absolute position information is affected by changes in scale, rotation, or viewpoint. However, the sorting method in proximity matrix can overcome these problems. In section 4, we perform the comparative experiments for the two matrices. Both theoretical analysis and experimental results prove spatio-temporal proximity matrix works better in the action recognition framework.

2.2. Proximity Distribution Kernel

After we have a global representation, spatio-temporal proximity matrix H, of each video sequence, we need to introduce a method to compare them. We define the Spatio-Temporal Proximity Distribution Kernel (ST-PDK) between two videos H_A and H_B as follows:

$$K(A,B)=K(H_A,H_B)=\sum_{i=1}^{K}\sum_{j=1}^{K}\min(H_A(i,j)-H_B(i,j)) \quad (4)$$

3. RECOGNITION BASED ON ST-PDK

In this section, we describe how to use the spatio-temporal proximity distribution for modeling human action classes. Then we adopt the SVM algorithm [16] for human action recognition.

For each video, we calculate a spatio-temporal proximity matrix for the detected features to describe the video. On one hand, the spatio-temporal proximity matrix captures the distribution of the video words which encodes the appearance of action. On the other hand, the proximity matrix characterizes the geometric consistency among actions in the same category. Hence the spatio-temporal proximity matrix encodes both appearance and the geometric relationship of human action class. Then, the similarity of proximity matrices is calculated as described in the section 2.

Finally, we incorporate ST-PDK into SVM classifier. It is clear that *K(A,B)* is a positive semi-definite kernel, and therefore it satisfies the Mercer's condition and is directly incorporated into the kernel function of the SVM classifier.

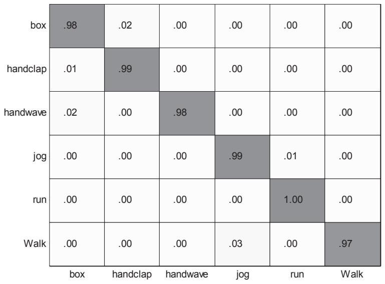
4. EXPERIMENTS

In this section we assess the performance of our recognition algorithm on KTH dataset (see Fig.1). This dataset is very challenging as it contains six types of human actions performed by 25 subjects in four different scenarios. Note that our recognition system directly manipulates the unsegmented input image sequences and does not require any preprocessing or object tracking [19]. In contrast, there is a common limitation in [12, 13, 15]: a figure centric spatio-temporal volume or silhouette for each person must be specified and adjusted with a fixed size in advance.

We perform leave-one-out cross-validation to make the performance evaluation. Moreover, there is no overlap between the training set and testing set. Specifically，the videos of the first two persons are used to produce the vocabulary of visual words.In each run, 24 actors' videos are used as the training set and the remaining one person's videos as the testing set. The final results are the average of 25 times runs.

We assess the performance of our recognition algorithm by reporting the results for a set of experiments. They show that: i) the approach based on ST-PDK outperforms the ordinary BOVW approach which is based on the distribution of video words; ii) the approach based on spatio-temporal proximity matrix outperforms the one based on spatio-temporal co-occurrence matrix; iii) different sizes of the vocabulary lead to different performances, but the approach based on ST-PDK achieves the best results; iv) we explore performance dependency with respect to different *r* in proximity matrix.

Fig.2 shows the confusion matrix of our approach on KTH dataset. Each row of the confusion matrix corresponds to the ground truth class, and each column corresponds to the assigned cluster. It shows that the "hand" related actions ("boxing", "handclapping", and "handwaving") are a little confused with each other, the "foot" related actions ("jogging", "running", and "waking") are a little confused with each other, but the "hand" related actions are totally differentiated from the "foot" related ones. Therefore, Fig.2 shows that ST-PDK is able to distinguish the "hand" related actions and the "foot" related ones due to its effectiveness in characterizing the geometric.



	box	handclap	handwave	jog	run	Walk
box	.98	.02	.00	.00	.00	.00
handclap	.01	.99	.00	.00	.00	.00
handwave	.02	.00	.98	.00	.00	.00
jog	.00	.00	.00	.99	.01	.00
run	.00	.00	.00	.00	1.00	.00
Walk	.00	.00	.00	.03	.00	.97

Fig. 2. The confusion matrix of our approach on KTH dataset.

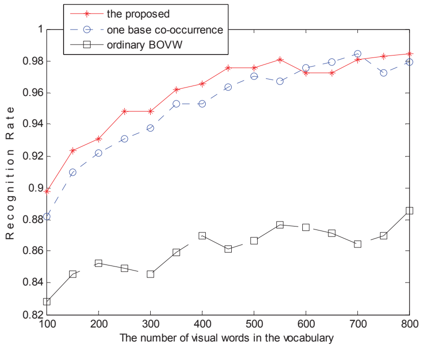


Fig.3. Recognition accuracy obtained by the three approaches vs. vocabulary size.

Further, we compare our ST-PDK based approach with two other approaches: the ordinary BOVW approach which is usually based on the distribution of video words, and the approach based on spatio-temporal co-occurrence matrix. Specifically, the ordinary BOVW approach employs a histogram of visual words to represent each video, uses χ^2 distance as the similarity of histograms, and then incorporates χ^2 distance into radial basis function as the kernel of SVM classifier. The final SVM kernel of the ordinary BOVW approach is formulated as follows:

$$K(A,B) = e^{-\frac{(\chi_{AB}^2)^2}{\sigma^2}} = e^{-\frac{1}{\sigma^2}\left(\sum_{j=1}^{K}\frac{(h_A(j)-h_B(j))^2}{h_A(j)+h_B(j)}\right)^2} \tag{5}$$

where h_A and h_B are the histograms of video *A* and *B* respectively, and σ is set to *2* in experiments by experience. For the approach based on co-occurrence matrix, we adopt the same experimental configurations with our approach except for using the spatio-temporal co-occurrence matrix instead of the spatio-temporal proximity matrix. Fig.3 draws the recognition accuracy curve of the three approaches vs. the vocabulary size *K*. it is shown that our approach gains the highest recognition accuracy in most cases. For *K*=[100, 150, 200,…, 800], our approach achieves 96% average recognition accuracy, which is 9.86% higher than the ordinary BOVW approach, and 0.8% higher than the co-occurrence matrix based approach. These experiments validate that our approach improves the recognition

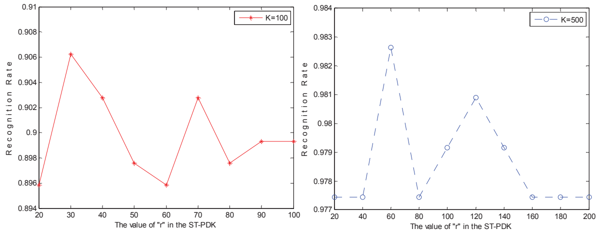


Fig.4. Recognition accuracy obtained by our approach vs. *r* of ST-PDK under two different vocabulary size *K*=[100,500].

Table 1. Comparison of different start-of-the-art methods.

Methods	Recognition Accuracy(%)
Schuldt et al. [2]	71.72
Dollár et al.[7]	81.17
Niebles et al. [4]	81.50
Jiang et al. [18]	84.40
Savarese et al.[5]	86.83
Kim et al. [17]	95.33
Our approach	**98.44**

accuracy of the ordinary BOVW approach by considering the spatio-temporal relationship of local features, and spatio-temporal proximity matrix is better than spatio-temporal co-occurrence matrix to characterize geometric information.

We assess the performances with respect to different *r*-th nearest neighbors in ST-PDK as illustrated in Fig. 4. The left one and the right one draw the recognition accuracy curve of our approaches vs. *r* under vocabulary size *K*=100 and *K*=500 respectively. The recognition accuracy fluctuates from 89.6% to 90.6% in the left one, and from 97.7 to 98.3 in the right one. The dependency of the recognition accuracy on the vocabulary size is not very significant. Thus, we set *r* to 60 in other experiments.

Table 1 compares the performances of our method with other recently developed methods in the literature. Up to our knowledge, our method achieves the best results compared with the current state-of-the art methods. Experiments show that ST-PDK indeed can improve the accuracy in human action recognition.

5. CONCLUSION

In this paper, we have developed a framework for recognizing low-level actions, such as walking, running, or handclapping, from input video sequences. In our recognition framework, the spatio-temporal proximity matrix of the local features is constructed to compensate for the geometrical unconstrained defect of BOVW. Experiments on the KTH dataset have proved the effectiveness and robustness of the proposed framework.

6. ACKNOWLEDGMENT

This work is partly supported by NSFC (Grant No. 60825204, 60672040, 60705003) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453, 2009AA01- Z318).

7. REFERENCES

[1] J. Liu, S. Ali, and M. Shah, “Recognizing Human Actions Using Multiple Features,” In CVPR, 2008.

[2] I. Laptev, and B. Caputo, “Recognizing Human Actions: A Local SVM Approach,” In *ICPR*, pp. 32- 36, 2004.

[3] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” In *CVPR*, 2008.

[4] J. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised Learning of Human Action Categories Using Spatial Temporal Words,” *IJCV*, pp. 299–318, 2008.

[5] S. Savarese, A. DelPozo, J. C. Niebles, Li Fei-Fei, “Spatial-Temporal correlatons for unsupervised action classification,” In *CVPR* 2008.

[6] F. Perronnin. “Universal and Adapted Vocabularies for Generic Visual Categorization,” *PAMI*, 30(7): 1243-1256, 2008.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior Recognition Via Sparse spatiotemporal Features,” In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65-72, 2005.

[8] J. G. Li, W. X. Wu, T. Wang, Y. M. Zhang, “One step beyond histograms Image representation using Markov stationary features,” In *CVPR*, pp.1-8, 2008.

[9] B. B. Ni, S. C. Yan, and A. Kassim, “Directed Markov Stationary Features for visual classification,” In *ICASSP*, pp. 825-828, 2009.

[10] H. B. Ling, S. Soato, “Proximity Distribution Kernels for Geometric Context in Category Recognition,” In *ICCV*, 2007.

[11] D. Q. Zhang and Z. H. Zhou, “(2D)2PCA: 2-Directional 2-Dimensional PCA for Efficient Face Representation and Recognition,” *Pattern Recognition*, 39(7): 1396-1400,2006.

[12] A. Fathi, and G. Mori, “Action Recognition by Learning Mid-level Motion Features,” In *CVPR*, 2008.

[13] K. Jia, and D. Yeung, “Human Action Recognition using Local Spatio-Temporal Discriminant Embedding,” In *CVPR*, 2008.

[14] C. F. Yuan, W. M. Hu, X. Li, S. Maybank, G. Luo, “Human action recognition under Log-Euclidean Riemannian metric”, In *ACCV* 2009.

[15] L. Wang, and D. Suter, “Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model,” In *CVPR*, 2007.

[16] C. Chang and C. Lin. *LIBSVM: a library for SVMs*, 2001.

[17] T. K. Kim, S. F. Wong, and R. Cipolla, “Tensor canonical correlation analysis for action classification,” In *CVPR*, 2007.

[18] H. Jiang, M. Drew, and Z. Li, “Successive convex matching for action detection,” In CVPR 2006.

[19] X.Q. Zhang, W. Hu, S. Maybank, and X. Li, “Graph based discriminative learning for robust and efficient object tracking”, In *ICCV* 2007.