Human Action Recognition using Pyramid Vocabulary Tree

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

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
{cfyuan, lixi, wmhu}@nlpr.ia.ac.cn
[2] School of Computer Science, University of Adelaide, SA 5005, Australia
wang.hanzi@gmail.com

Abstract. The bag-of-visual-words (BOVW) approaches are widely used in human action recognition. Usually, large vocabulary size of the BOVW is more discriminative for inter-class action classification while small one is more robust to noise and thus tolerant to the intra-class invariance. In this pape, we propose a pyramid vocabulary tree to model local spatio-temporal features, which can characterize the inter-class difference and also allow intra-class variance. Moreover, since BOVW is geometrically unconstrained, we further consider the spatio-temporal information of local features and propose a sparse spatio-temporal pyramid matching kernel (termed as SST-PMK) to compute the similarity measures between video sequences. SST-PMK satisfies the Mercer's condition and therefore is readily integrated into SVM to perform action recognition. Experimental results on the Weizmann datasets show that both the pyramid vocabulary tree and the SST-PMK lead to a significant improvement in human action recognition.

Keywords: Action recognition, Bag-of-visual-words (BOVW), Pyramid matching kernel (PMK)

1 Introduction

Human action recognition has been received more and more attentions due to its crucial values in smart surveillance, human-computer interface, video indexing and browsing, automatic analysis of sports events, and virtual reality. However, there exist many difficulties with human action recognition, including occlusion, illumination changes, as well as geometric variations in scale, rotation, and viewpoint.

In general, the action recognition approaches can be roughly classified as the template-based and the appearance-based approaches [1]. For the template-based approaches, there exist two sorts of templates. The first sort of templates directly use several key frames or segmented patches of the input videos, as described in [6, 8]. The second sort of templates are obtained by linear or nonlinear transformation of the input videos. For example, Rodriguez et al. [9] combine a sequence of training images into a single composite template by a MACH filter. For the appearance-based approaches, local features or global (or large-scale) features are employed to represent the videos. Generally, local spatio-temporal features are more robust to noise, occlusion and action variation than large-scale features.

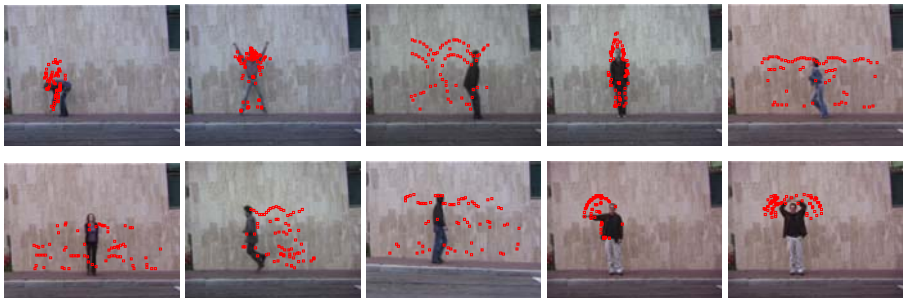

Fig. 1. Interest points localization of ten action video sequences in Weizmann dataset. Each red point corresponds to a video patch associated with a detected interest point. One key frame is shown for each video and all interest points detected in that video are overlapped on the key frame.

Recently, several state-of-the-art action recognition approaches [2, 3, 4, 5, 17, 19] use the BOVW to exploit local spatio-temporal features. Typically, these approaches firstly generate a vocabulary of visual words and then characterize videos with the histograms of visual word counts. It is obvious that the vocabulary plays a decisive role in the process of action recognition. A good vocabulary should not only discriminate the inter-class invariance but also tolerant the intra-class invariance of objects or actions. It is common to choose an appropriately large vocabulary size [4, 10]. However, the large size of vocabulary may introduce a sparse histogram for each video, yield more noise and reduce the discriminability of vocabulary. On the other side, if the vocabulary size is small, it may cause over-clustering and high intra-class distortion. Motivated by these observations, we propose a novel architecture of vocabulary – *the pyramid vocabulary tree* which combines the vocabularies of different sizes and exploits a larger and more discriminative vocabulary efficiently. In addition, it is very fast to project new features on the tree. With pyramid vocabulary tree, video sequences are hierarchically represented as the multi-resolution histograms of the vocabulary tree.

Moreover, it is well known that the BOVW approaches are geometrically unconstrained. Therefore, there are many algorithms intending to combine the geometrical information with BOVW. Some approaches [13, 15] uniformly divide the 3D space into the spatio-temporal grids and then compute the histogram of local features in each grid. However, in the human action videos, the interest points are usually detected in some local regions while most other regions contain no interest points (as illustrated by Fig.1). Inspired by this observation, we cluster the interest points in the spatio-temporal space, which forms several cluster centers. At each cluster center we compute the histogram of the local features. Based on the spatio-temporal cluster centers, we propose a sparse spatio-temporal pyramid matching kernel (called SST-PMK) to measure the similarities between video sequences. In SST-PMK, the histogram used for representing the video is more compact and robust than that in [13, 15]. Therefore the distance computed by SST-PMK is more reliable. Besides, the SST-PMK satisfies the Mercer's condition and can be directly used as the SVM kernel to perform action recognition.

In general, we propose a novel framework based on the sparse spatio-temporal representation of the pyramid vocabulary tree for action recognition. The pyramid tree is built to model the local features, and also prepares a hierarchical structure for computing SST-PMK. Moreover, SST-PMK effectively integrates the distances obtained from all levels of the pyramid vocabulary tree to compute the similarities between video sequences with a very fast speed.

The remainder of the paper is organized as follows. Section 2 describes how to generate the pyramid vocabulary tree. Section 3 introduces SST-PMK and then combines it with the SVM classifier. Section 4 reports experimental results. Section 5 concludes the paper.

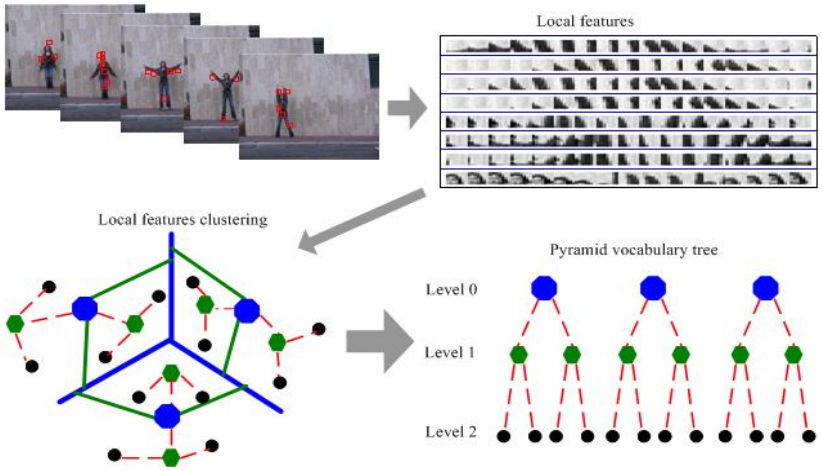

Fig. 2. The building process of the proposed pyramid vocabulary tree.

2 Pyramid Vocabulary Tree

The Pyramid vocabulary tree is built by hierarchically clustering a large set of training descriptor vectors. The building process of the pyramid vocabulary tree is illustrated in Fig. 2. First of all, the training descriptor vectors are clustered into k visual words to build the coarsest level 0 (i.e. the conventional BOVW). Subsequently, we split each visual word at the coarsest level 0 into two ones, resulting in a finer vocabulary level. In this case, the vocabulary tree grows in a hierarchical coarse-to-fine manner. Meanwhile, the number of its leaf nodes increases in an exponential manner.

In the following sections, we briefly describe the generation of BOVW and the building of the pyramid vocabulary tree in details.

2.1 The generation of BOVW

A large set of local features are used in the unsupervised training of the tree. Capturing local features includes two relatively independent steps: detecting cuboids

and describing cuboids. In recent years, a number of detectors and descriptors have been proposed for human action recognition. All can be used in our recognition framework. In this paper, we employ the Dollár et al.'s detector [7] to detect cuboids at every frame of each video and use the PCA-SIFT descriptor [14] to describe the detected cuboids. Dollár et al. [7] detector improves the 3D Harris detector by applying Gabor filtering to the temporal domain. The outputs of the detector are the location, the scale, and the dominant orientation of each interest point. We extract a cuboid at a given scale centered at every interest point with a size which is s times of its scale (s is set to be 6 in this paper). Then, PCA-SIFT descriptor applies Principal Components Analysis (PCA) to the normalized gradient vector which is formed by flattening the horizontal and vertical gradients of all the points in the cuboid.

Subsequently, a K-means clustering process is run on the obtained PCA-SIFT features. As a result, k cluster centers are treated as k visual words at Level 0. Other clustering methods, such as spectral clustering [21] or Maximization of Mutual Information (MMI) [22], can also be two alternatives instead of the K-means clustering.

2.2 The pyramid vocabulary tree

After building the 0^{th} level of the tree, the training features are partitioned into k groups, where each group consists of the features closest to a particular visual word. Then the training features of each group are clustered into two new visual words at a new level. Therefore each visual word at 0^{th} level is split into two new visual words at level 1. This splitting is reasonable because the visual words at level 0 are highly compact after clustering. In this way, the tree grows till the maximum number of levels L is reached. The vocabulary size of each level is doubled than its upper level.

In the online phase, each new PCA-SIFT feature is compared to k candidate cluster centers at level 0 and assigned to the nearest words. Then the result is propagated to the next level in order that we only need to compare the descriptor vector to the 2 children cluster centers and choose the closest one. Level by level, the new feature is projected to the tree very fast. Furthermore, in the computational complexity aspect, the quantization of new PCA-SIFT features requires $k+2L$ dot products in our approach. However, it needs $2^L k$ dot products for the conventional BOVW in a non-hierarchical manner with the same vocabulary size at the L^{th} level.

3 SVM classification based on SST-PMK

With the pyramid vocabulary tree, each video can be represented as a multi-level visual word histogram. To effectively measure the similarity of two visual word histograms, we present a sparse spatio-temporal pyramid matching kernel (called SST-PMK) in this section. Moreover, SST-PMK can serve as a kernel for SVM classification.

3.1 The sparse spatio-temporal pyramid matching kernel (SST-PMK)

The pyramid matching kernel (PMK) proposed by Grauman and Darrell [11] is an effective kernel to measure the similarity of two multi-resolution histograms and it has been successfully applied to object recognition. However, one potential problem with the PMK [11] is that it does not consider the spatio-temporal information. From Fig.1, it can be seen the geometrical distribution of interest points is regularly varying among different action classes, and thus spatio-temporal information is very helpful for improving the action recognition accuracy. Therefore, we take into account the spatio-temporal information of interest points while computing PMK. This is the contribution of our SST-PMK.

The other observation in Fig.1 is that interest points are not uniformly distributed in the image and some regions contain no interest points. Without considering this observation, the SPM [13] uniformly partitions the whole image into 2D grids in the spatial space (i.e., the image coordinate) and the STPM [15] uniformly partitions the whole video into 3D grids in the spatial and temporal space. These two methods do not effectively assign grids, which leads to a large number of grids and some of the grids do not contain any interest points. Moreover, both SPM and STPM require a preprocessing step for normalizing the size of images or videos. In contrast, the grids obtained by SST-PMK are sparse and discriminative, without normalizing videos beforehand. Fig. 3 shows the hierarchical structure of SST-PMK. The following lists the specific procedure of constructing the SST-PMK.

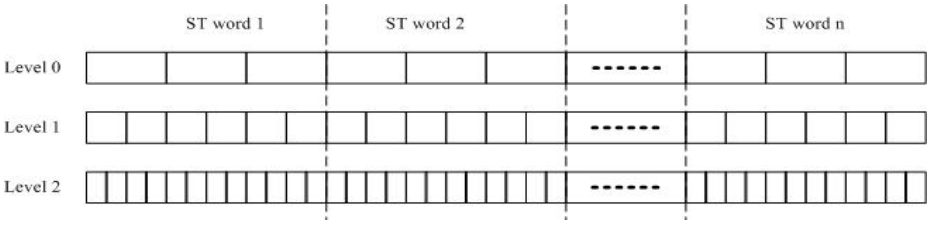

Fig.3. The hierarchical structure of SST-PMK for each video. The geometrical information of interest points is combined with the pyramid vocabulary tree to represent the videos.

At first, the spatio-temporal vectors of interest points are clustered to produce spatio-temporal words (i.e. ST word i in Fig. 3 $1 \le i \le n$). The 3-D data set formed by these vectors is divided into several subsets. The ST words are derived from the center of subsets.

Then, for each video, we compute the histograms of its descriptor vectors (i.e. PCA-SIFT features) at each ST word and each level. And then we concatenate the obtained histograms into a vector $H = [H_1, \ldots, H_L]$, where H_l represents the histogram at level l. And $H_l = [h_{l\text{-}ST\,1}, \ldots, h_{l\text{-}ST\,n}]$, where $h_{l-ST\,i}$ is the histogram for ST word i at level l. That is, we build a hierarchical structure as Fig.3 for each video and represent the video as a histogram vector.

Given the corresponding histogram vectors X and Y of two videos, the SST-PMK computes a weighted histogram intersection in the hierarchical structure as illustrated in Fig. 3. At each level l, the histogram intersection is defined as the sum of the minimal value at each bin:

$$
\begin{aligned}
I(X_l, Y_l) &= \sum_m \min(X_l(m) - Y_l(m)) \\
&= \sum_{i=1}^{2^l k} \min(x_{l-ST1}(i) - y_{l-ST1}(i)) + \cdots + \sum_{i=1}^{2^l k} \min(x_{l-STn}(i) - y_{l-STn}(i)) \\
&= \sum_{j=1}^{n} \sum_{i=1}^{2^l k} \min(x_{l-STj}(i) - y_{l-STj}(i))
\end{aligned} \tag{1}
$$

where x_{l-STj} is an element of X and represents the histogram of the video for ST word j at the level l, and $x_{l-STj}(i)$ denotes the count of the i^{th} bin of x_{l-STj}. The number of the newly matched pairs N_l induced at level l is the difference between successive levels' histogram intersections:

$$
N_l = I(X_l, Y_l) - I(X_{l+1}, Y_{l+1}) \tag{2}
$$

Because level L is the finest level, we compute the number of matches N_l from level L to level 0 just opposite to the building process of the pyramid vocabulary tree. The resulting kernel K is obtained by the weighted sum over the number of matches N_l occurred at each level, and the weight associated with level l is set to (2^{l-L}):

$$
\begin{aligned}
K(X,Y) &= \sum_{l=0}^{L} \frac{1}{2^{L-l}} (I(X_l, Y_l) - I(X_{l+1}, Y_{l+1})) \\
&= \sum_{l=0}^{L-1} \frac{1}{2^{l+1}} I(X_{L-l}, Y_{L-l}) + \frac{1}{2^L} I(X_0, Y_0) \\
&= \sum_{j=1}^{n} \left(\sum_{l=0}^{L-1} \left(\frac{1}{2^{l+1}} \sum_{i=1}^{2^{L-l} k} \min(x_{(L-l)-STj}(i) - y_{(L-l)-STj}(i)) \right) \right. \\
&\qquad \left. + \frac{1}{2^L} \sum_{i=1}^{k} \min(x_{0-STj}(i) - y_{0-STj}(i)) \right)
\end{aligned} \tag{3}
$$

where $X_{L+1} = Y_{L+1} = 0$.

The SST-PMK effectively combined each level in the hierarchical structure. The newly matched pairs at coarser level, which are not matched ones at its finer level, are also considered in the SST-PMK. This corresponds to some cases in action recognition, such as the same class actions performed by different persons, and the same class actions performed by one person at many times. If these intra-class actions are not regarded as match at fine level, they are still able to be treated as match at coarser level. Therefore, according to the pyramid tree and SST-PMK, our approach can overcome the variations between intra-class objects or actions.

3.2 SVM classification

We adopt the algorithm in [16] to train SVM for human action recognition. From equation (3), we obtain the following equation:

$$
K(X,Y) = \sum_{j=1}^{n} K_{\Delta}(X_{STj}, Y_{STj}) \tag{4}
$$

$$
\begin{aligned}
K_{\Delta}(X_{STj}, Y_{STj}) &= \sum_{l=0}^{L-1} \left(\frac{1}{2^{l+1}} \sum_{i=1}^{2^{L-l} k} \min(x_{(L-l)-STj}(i) - y_{(L-l)-STj}(i)) \right) \\
&\quad + \frac{1}{2^L} \sum_{i=1}^{k} \min(x_{0-STj}(i) - y_{0-STj}(i))
\end{aligned} \tag{5}
$$

$K_{\Delta}(X_{STj}, Y_{STj})$ is actually a pyramid matching kernel (PMK) [11]. In [11] it is proved that PMK is a Mercer kernel and a positive semi-definite kernel. Given that Mercer

kernels are closed under addition, equation (4) shows that SST-PMK is a Mercer kernel. Therefore, SST-PMK distance between videos is directly incorporated into the kernel function of the SVM classifier.

4 Experiments

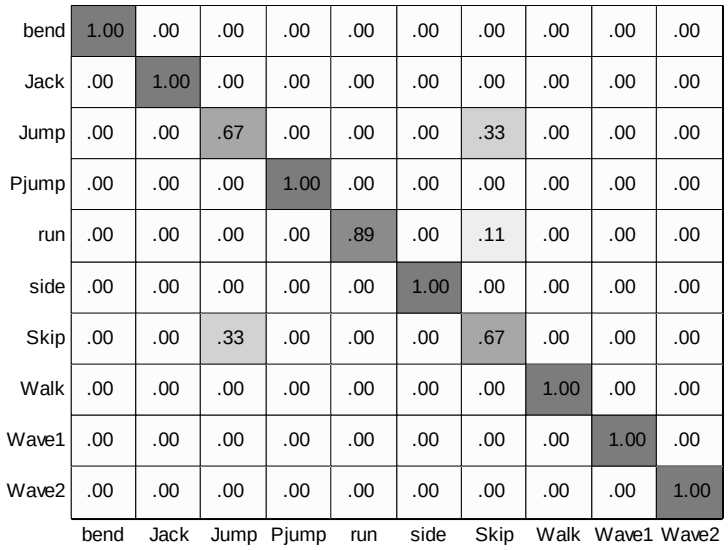

Fig. 4. The confusion matrix of our approach on the Weizmann action dataset.

The proposed action recognition approach directly manipulates the unsegmented input image sequences, which aims to recognize low-level actions such as walking, running, and hand clapping. Note that our recognition system does not require any preprocessing step. In contrast, there is a common limitation in [12, 18, 20]: a figure centric spatio-temporal volume or silhouette for each person must be specified and adjusted with a fixed size in advance. However, object segmentation and tracking is hard to implement in itself.

We test our approach on the Weizmann dataset [23]. The Weizmann human action dataset contains 10 different actions including Walking, Running, Jumping, Galloping sideways, Bending, One-hand waving, Two-hands waving, Jumping in place, Jumping Jack and Skipping. One representative frame from each action category is shown in Fig.1. There are 93 samples in total. The resolution of the videos is 320x240 pixels and the frame rate is 15fps.

We perform the leave-one-out cross-validation to evaluate the competing algorithms. The red line is obtained by the proposed approach, the blue one is the ordinary BOVW approach, and the black one is the PMK approach without considering the spatio-temporal information.

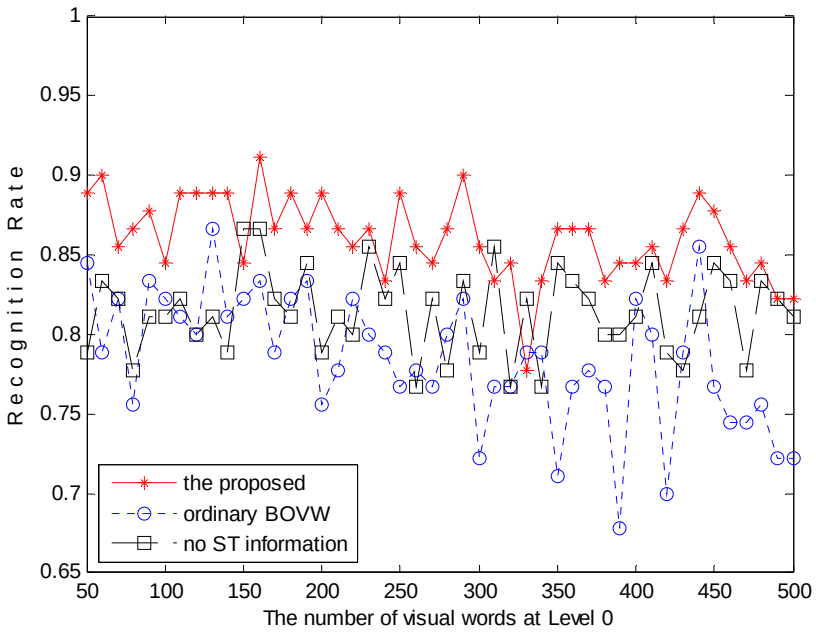

Fig.5. Recognition accuracy obtained by the three approaches vs. vocabulary size in Level 0.

In all experiments we use the videos of the first five persons to learn the bag of visual words. In each run, the videos of 8 actors are used as the training set and the remaining videos of one person are used as the testing set. There is no overlap between the training set and the testing set. We run the algorithms 9 times and report the average results.

In our approach, the three-level pyramid vocabulary tree is used to model local features. The number of visual words is set to 160 at the coarsest level (i.e. level 0) and 640 at the finest level (i.e. level 2). The geometrical information of the interest points is clustered into 10 centers. We use SST-PMK as the SVM kernel. Fig. 4 shows the confusion matrix of our approach on the Weizmann dataset. Each row of the confusion matrix corresponds to the ground truth class, and each column corresponds to the assigned cluster. It shows that our approach works much better on the actions with large movements, but it does not achieve desirable results for the actions of small difference. The recognition accuracy for the actions with large movements is 100%, such as “bend”, “Jack”, “Pjump”, “side”, “walk”, “wave1”, and “wave2”. The actions “Jump”, “Run”, and “Skip” are similar to each other, and thus may be a little confused with each other.

4.1 The comparison of three approaches

In order to demonstrate the advantages of the pyramid vocabulary tree and the

proposed SST-PMK, we compare two other approaches with our approach. In the first approach, we use only one vocabulary (i.e. conventional BOVW) and the remaining settings are all the same as our approach. Since there is only one level, the SST-PMK degenerates to the sum of the two histogram intersection:

$$I(X,Y)=\sum_{i=1}^{nk_L}\min(X(i),Y(i)) \quad (6)$$

where n is the number of ST words, and k_L is equal to the vocabulary size of level L in our approach. Therefore in the first approach, equation (6) is used as the kernel of SVM classification. For the second approach, we do not consider the geometrical information, i.e., PMK is used for SVM classification. Moreover, we run the three approaches using different vocabulary sizes. Fig.5 draws the recognition accuracy curve of the three approaches vs. the vocabulary size k at level 0. Fig.5 shows that our approach gains the highest recognition accuracy at most cases. For k=[50, 60, …, 500], our approach is on average 7.63% higher than the first approach, and 4.66% higher than the second approach. It demonstrates that both the pyramid vocabulary and the geometrical information of the interest points are helpful for the action recognition.

4.2 Kernel comparison of SVM

Table 1. Comparisons between the proposed SST-PMK and the four popular kernels for SVM classifier.

	Linear	Polynomial	RBF	Sigmoid	**SST-PMK**
Bend	**1**	0.6667	**1**	**1**	**1**
Jack	**1**	0.4444	**1**	**1**	**1**
Jump	**0.8889**	0.4444	0.6667	0.5556	0.6667
Pjump	0.8889	0.4444	**1**	**1**	**1**
Run	0.6667	0.5556	0.8889	0.8889	**0.8889**
Side	**1**	0.1111	**1**	**1**	**1**
Skip	**0.6667**	0.4444	**0.6667**	0.4444	**0.6667**
Walk	0.7778	0.2222	0.8889	**1**	**1**
Wave1	**1**	0.2222	**1**	**1**	**1**
Wave2	**1**	0.6667	**1**	**1**	**1**
Average	0.8889	0.4222	0.9111	0.8889	**0.9222**

We also compare the proposed SST-PMK with other four popular kernels used in SVM: linear kernel $x'*y$, polynomial kernel $(g*x'*y)^3$, radial basis function (RBF) $exp(-g||x-y||^2)$, and sigmoid kernel $tanh(g*x'*y)$. The same experimental configurations are used for all five kernels. Moreover, in the SVM classifier [16], C-Support Vector Classification (C-SVC) is employed and two kernel parameters (c and g) are considered. Different kernel parameters are used to estimate the recognition accuracy:

$$c=[2^{-5},2^{-4},\cdots,2^{25}],\quad g=[2^{-15},2^{-14},\cdots,2^{3}]$$

More specifically, since the linear kernel and SST-PMK just have one parameter c,

we try 31 different c values and report the best results. For the other three kernels (polynomial kernel, RBF, and sigmoid kernel) have two parameters c and g, we try 31×19=589 combinations. Table 1 shows the experimental results using the five kernels based approaches. Polynomial kernel based approach achieves the worst results, and the average accuracy of other three kernels (Linear kernel, Sigmoid kernel, and RBF) based approaches is a little lower than ours. Our approach achieves the best recognition performances, and it outperforms the other four kernels for nine actions of ten.

5 Conclusion

In this paper, we develop a novel framework which can recognize low-level actions such as walking, running, and hand clapping from unsegmented video sequences. This paper has the following two contributions. First, to the best of our knowledge, the vocabulary is built into pyramid tree topology in human action recognition for the first time. Second, we propose SST-PMK, which takes advantages of geometrical information of local features, to compute the similarities between video sequences. SST-PMK improves PMK by clustering the spatio-temporal information of interest points. Experiments show the effectiveness and robustness of the proposed approach.

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).

References

1. J.K. Aggarwal, and S. Park. Human motion: modeling and recognition of actions and interactions. In *Second International Symposium on 3D Data Processing, Visualization and Transmission*, pp. 640-647, Sep 6 – 9, 2004.
2. C. Schuldt, 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. K. Yan, R. Sukthankar, and M. Hebert. Efficient Visual Event Detection using Volumetric Features. In *ICCV*, pp. 166- 173, 2005.
6. D. Weinland, and E. Boyer.Action Recognition using Exemplar-based Embedding. In *CVPR*, 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. F. Lv, and R. Nebatia. Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching. In CVPR, 2007.
9. M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition. In CVPR, 2008.
10. B. Fulkerson and A. Vedaldi, and S.Soatto. Localizing Objects With Smart Dictionaries. in *Proceedings of the European Conference on Computer Vision* (*ECCV*), 2008.
11. K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. In Proc. *ICCV*, 2005.
12. A. Fathi, and G. Mori. Action Recognition by Learning Mid-level Motion Features. In *CVPR*, 2008.
13. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pp. 2169-2178, 2006.
14. K. Yan, and R. Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *CVPR*, pp. 506-513, 2004.
15. J. Choi, W. J. Jeon, and S. C. lee. Spatio-Temporal Pyramid Matching for Sports Videos. In *Proceedings of ACM International Conference on Multimedia Information Retrieval (MIR)*, 2008.
16. C. Chang and C. Lin. *LIBSVM: a library for SVMs*, 2001.
17. J. Liu, S. Ali, and M. Shah. Recognizing Human Actions Using Multiple Features. In *CVPR*, 2008.
18. K. Jia, and D. Yeung. Human Action Recognition using Local Spatio-Temporal Discriminant Embedding. In *CVPR*, 2008.
19. F. Perronnin. Universal and Adapted Vocabularies for Generic Visual Categorization. *PAMI*, 30(7): 1243-1256, 2008.
20. L. Wang, and D. Suter. Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. In *CVPR*, 2007.
21. Y. Wang, H. Jiang, M.S. Drew, Z. Li and G. Mori. Unsupervised Discovery of Action Classes. In *CVPR*, pp. 1654- 1661, 2006.
22. J. Liu, and M. Shah. Learning Human Actions via Information Maximazation. In *CVPR*, 2008.
23. www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html.