Bag of Spatio-temporal Synonym Sets for Human Action Recognition

Lin Pang[1,2], Juan Cao[1], Junbo Guo[1], Shouxun Lin[1], and Yan Song[1,2]

[1] Laboratory of Advanced Computing Research, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] Graduate University of Chinese Academy of Sciences, Beijing, China
{panglin,caojuan,guojunbo,sxlin,songyan}@ict.ac.cn

Abstract. Recently, bag of spatio-temporal local features based methods have received significant attention in human action recognition. However, it remains a big challenge to overcome intra-class variations in cases of viewpoint, geometric and illumination variance. In this paper we present Bag of Spatio-temporal Synonym Sets (ST-SynSets) to represent human actions, which can partially bridge the semantic gap between visual appearances and category semantics. Firstly, it re-clusters the original visual words into a higher level ST-SynSet based on the distribution consistency among different action categories using Information Bottleneck clustering method. Secondly, it adaptively learns a distance metric with both the visual and semantic constraints for ST-SynSets projection. Experiments and comparison with state-of-art methods show the effectiveness and robustness of the proposed method for human action recognition, especially in multiple viewpoints and illumination conditions.

Keywords: Action Recognition, Spatio-temporal Synonym Sets, Metric Learning.

1 Introduction

Human action recognition is an important technique for multiple real-world applications, such as video surveillance, human-computer interaction, and event-based video retrieval. However, it still remains a challenging task due to the cluttered background, camera motion, occlusion and viewpoint variance, etc.

There are two kinds of traditional methods for action recognition. One is global model based method. For example, Bobick et al. in [1] use motion history images to recognize actions and Blank et al. in [2] represent actions by describing the spatio-temporal shape of silhouettes. However, these methods rely on the restriction of contour tracking and background subtraction. Without prior foreground segmentation, Efros et al. in [3] correlate flow templates with videos and Shechtman et al. in [4] recognize actions by spatial-temporal volume correlation, but these methods are quite sensitive to scale, pose and illumination changes. The other is the local feature based method. Recently, the spatio-temporal local feature based method has been widely

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 422–432, 2010.

used for human behavior analysis in [5, 6, 7, 8, 9, 10]. By using a Bag-of-Words representation combined with machine learning techniques like Support Vector Machine [5, 7] and graphical models [6], it performs better than the global model based method, especially in the situation with cluttered background and severe occlusion.

In the above spatio-temporal local feature based method, quantizing the local features into visual words is one of the key problem. Among the majority works to date, the visual words are usually obtained by unsupervised clustering methods such as K-means. Niebles et al. in [6] learn the latent topics of the visual words using generative graphical models such as the probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA). To get a compact representation, Liu et al. in [10] propose an unsupervised method with Maximization Mutual Information principle to group visual words into video-words-clusters (VWC). Essentially, these clusters of visual words obtained in the unsupervised manner only capture the common visual patterns in the whole training set, thus may not have the most discriminative power in accordance with action category semantics. However, one of the most significant challenges in human action recognition is the different visual appearances of local features within the same action category in cases of viewpoint changes as well as geometric and illumination variance. Meanwhile, actions of different categories may share similar local appearances. An example of intra-class variation and inter-class similarity is shown in Fig.1. It is clear to see that local cuboids a and b with relatively different visual appearances caused by viewpoint change are from the same action category "Running", while cuboids b and c with higher visual similarity are from different action categories. Therefore, spatial-temporal visual vocabularies constructed only by unsupervised visual appearance similarity clustering are limited to handle this gap between visual appearances and category semantics, and the ambiguous vocabularies projection in the visual feature space may hurt the overall classification performance.

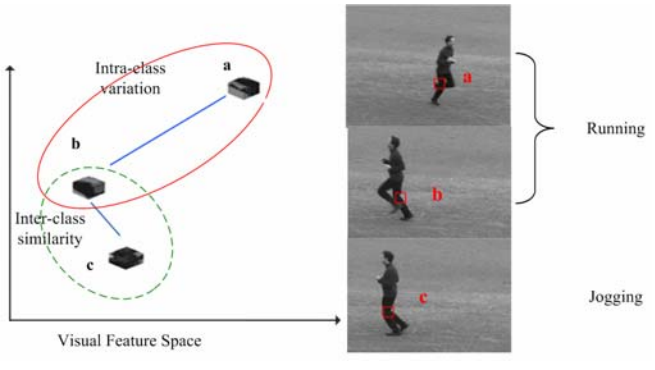


Fig. 1. Intra-class variation and Inter-class similarity of local cuboids in action recognition. Cuboids a and b are from two samples of "Running" with different viewpoints, and cuboid c is from sample of "Jogging".

To address the above issue, motivated by the research in object recognition [11, 12], we propose bag of ***Spatio-temporal Synonym Sets (ST-SynSets)*** for representation of actions to partially bridge the gap between visual appearances and category semantics, where ST-SynSet is a higher level cluster of semantic consistent visual words. The main idea is that though it is hard to measure the semantics of visual words directly, visual words from the same action category with variant visual appearances in cases of scale, viewpoint and illustration change could still keep the similar probability distributions among different categories, which in a way means the semantic similarity. In addition, since the topological proximity of visual words in the visual feature space can't ensure the semantic relevance, vocabulary projection just by visual nearest neighbor mapping lacks the concordance with category semantics. Thus we need to learn a new distance metric for visual vocabulary by integrating the visual and semantic similarity and transfer the visual words to the ST-SynSet space to suppress the errors caused by uncertainty of vocabulary projection. The ST-Synset is different from latent topic in pLSA and LDA in[6] for that it is not a result of a generative model. Without prior assumption of the distribution, the ST-SynSet is the result of a supervised data-mining process of compressing visual words via distributional clustering following the joint distribution of visual words and action categories.

The main contribution lies in two aspects:

First, we propose to cluster visual words which share similar category probability distributions to be ST-Synsets by the Information Bottleneck clustering method, and produce a compact and discriminative representation for actions.

Second, we propose to learn a new distance metric for visual vocabularies based on the synonymy constraints to get a more accurate ST-SynSet projection.

The remainder of the paper is organized as follows. Section 2 presents the proposed method in detail. Experimental results are shown in Section 3, and Section 4 concludes this paper.

2 Spatio-temporal Synonym Sets Based Action Recognition

Fig. 2 shows the flowchart of the proposed action recognition algorithm. First, we adopt the spatio-temporal interest points detector proposed by Dollar et al. in [5] for local feature extraction. This detector produces dense feature points and performs well on the action recognition task [6]. For each cuboid we compute gradient-based descriptor and apply PCA to reduce dimensionality. Then, initial visual vocabularies are constructed by clustering the extracted spatio-temporal local features with K-means algorithm. Second, Sequential Information Bottleneck (SIB) method is implemented to re-cluster the visual words into Spatio-temporal Synonym Sets and informative ST-SynSets are selected by the information score. Third, in order to get a reasonable ST-SynSet projection, we learn a new distance metric for visual vocabulary with the synonym constraints. Finally, we use "Bag of ST-SynSets", the histogram of Spatio-temporal Synonym Sets to represent each action instance, and use Support Vector Machine (SVM) for human action recognition.

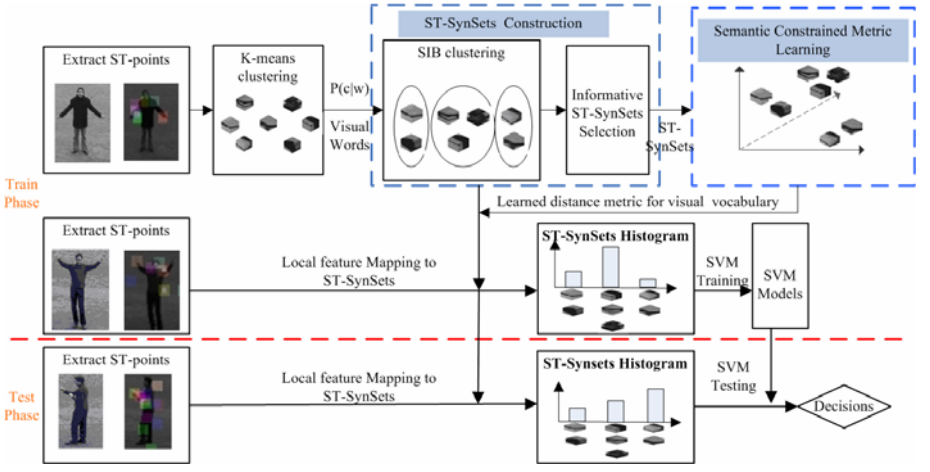


Fig. 2. Flowchart of the proposed action recognition algorithm

2.1 Spatio-temporal Synonym Sets Construction

In the state-of-art method of Bag-of-Words, an action instance is encoded as a histogram of visual words by unsupervised vector quantization of local features. However, action instances usually have significant intra-class variations because of the different attributes of performers (age, gender, clothes, and velocity) and especially different external conditions, such as viewpoints, scales, illuminations and so on .Therefore, the visual words with only the similarity of visual appearances become too primitive to effectively represent the characteristic of each category.

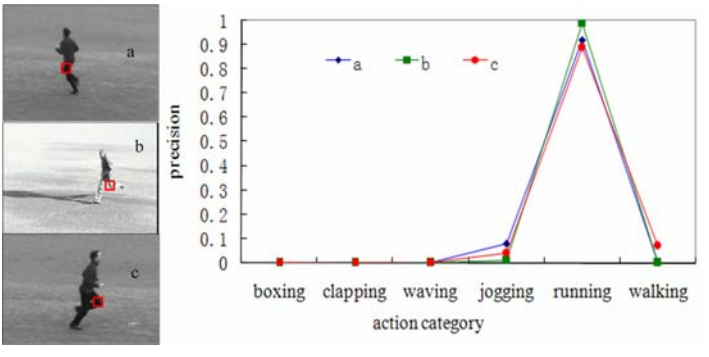


Fig. 3. Category probability distributions of three cuboids with different appearances extracted from "Running "actions in three different scenarios

Motivated by the object recognition method in [12], we find that local features highly correlated with the same human action category may vary in visual appearances under different circumstances but they keep the discriminative ability among different categories. Thus, we define $P(c_i \mid w)$ to measure the semantic inference

ability of visual word w attributed to particular category c_i. Due to the motion pattern heterogeneity of actions, a number of local features are intrinsic and highly indicative to certain action categories. For example in Fig. 3, three visually different cuboids from "Running " in three different scenarios with large viewpoint variances have similar category probability distributions $P\ (c \mid w)$ which peak around its belonging classes and denote the semantic similarity of the local cuboids.

Therefore, we define ***Spatio-temporal Synonym Set*** to be the cluster of semantic-consistent visual words which share the similar probabilistic distributions $P\ (c \mid w)$ among different categories. This higher level representation groups visual words with different visual appearances but similar discriminative power together and thus can partially handle the significant intra-class variations of local features and have more discriminative power to distinguish between action categories.

2.1.1 Distributional Clustering by Information Bottleneck Method

Based on the definition of ST-SynSet, we use the Information Bottleneck (IB) principle [13] which provides a reasonable solution for distributional clustering to cluster the initial visual words to a compact representation for the construction of ST-SynSets. In [10], Liu et al. use the Maximization Mutual Information principle to group visual words into video-words-clusters (VWC). In essence, the visual word clusters obtained in this unsupervised manner mean to capture the common visual patterns with similar distributions among all the video samples in the training set, thus may lack the discriminative power to distinguish different action categories. In our method, we get the most compact representation of visual words and meanwhile maintain as much discriminative information of the categories as possible by clustering the visual words with similar probabilistic distributions $P\ (c \mid w)$ among different categories using the IB principle.

Given the joint distribution $P(w, c)$ of each visual word w and action category c, the goal of IB principle is to construct the optimal compact representation of visual words set W, namely the ST-SynSets S, such that S preserves as much information of category set C as possible. The IB principle is formulated as the following Lagrangian optimization problem in Equation (1).

$$\max_{S} F(S) = I(S;C) - \beta I(W;S) \tag{1}$$

where $I(S;C)$ and $I(W;S)$ are the mutual information between S and C and between W and S respectively. And the mutual information $I(X; Y)$ is defined as

$$I(X;Y) = \sum_{x \in X,\, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

Equation (1) means to cluster the visual words into the most compact representation S with the least mutual information with visual words W through a compact bottleneck under the constraint that this compression maintains as much information of the category set C as possible.

In [13], Noam et al. propose a sequential IB clustering algorithm to solve the optimization problem in Equation (1) with the $\beta \to \infty$ limit to generate the hard partitions.

The algorithm starts from an initial random partition S_0 of W, and at each step we draw each w out of its current cluster $S(w)$ and choose for its new cluster by minimizing the score loss caused by merging w to every cluster s_i which is stated in Equation(3). Repeat this process until convergence. To avoid the algorithm being trapped in local optima, we repeat the above procedure for random initializations of S_0 to obtain n different solutions, from which we choose the one that maximize $F(s)$ in Equation(1).

$$d_F(w, s_i) = (P(w) + P(s_i)) \times JS(P(c|w), P(c|s_i)) \tag{3}$$

where c is the variable of category, s_i is the i-th cluster in the current partition S, and $JS(p,q)$ is the Jensen-Shannon divergence [14] which essentially measures the likelihood that the two sample distributions p and q originate from the most likely common source.

2.1.2 Informative ST-SynSet Selection

After distribution clustering, some of the ST-SynSets which have flat and non-salient category probability distributions are not discriminative and have more uncertainty in semantics. Therefore an effective ST-SynSet selection is necessary. We define the maximal mutual information between the ST-SynSet s and each category label c to measure the information score of s, which is formulated in Equation (4).

$$I(s) = \max_c (I(s,c)) = \max_c \sum_{w \in s, c \in C} p(w,c) \log \frac{p(w,c)}{p(w)p(c)} \tag{4}$$

We select the most significant ST-SynSets with the highest $I(s)$ and remove the others with lower $I(s)$. The local appearance samples of cuboids from the three most informative ST-SynSets and their corresponding category probability distributions are shown in Fig 4. From the spatio-temporal patches we can see that they are all signature motion parts of the corresponding action categories.

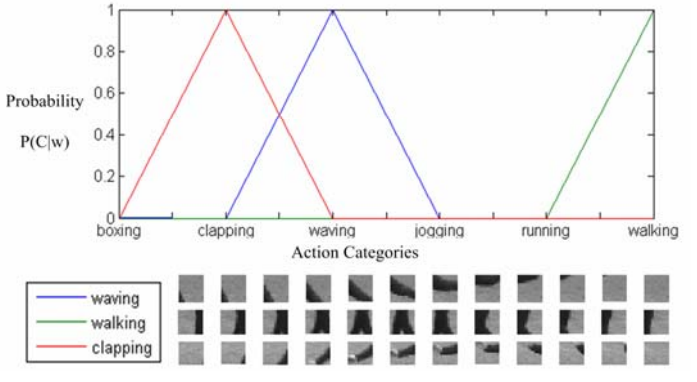


Fig. 4. Local appearance samples of cuboids from the top three informative ST-SynSets and their corresponding category probability distributions

2.2 Distance Metric Learning with Semantic Constraints

The basic mechanism of Bag-of-Words approach involves a step of mapping the local features to visual words according to the distances between the local features and the visual words. In the proposed approach, we need to map the local features to ST-SynSets. It is known that the visual words in the same ST-SynSet may have different visual appearances, while those with similar appearances may represent different category semantics. The inconsistency between semantic space and visual feature space causes the ambiguity and uncertainty of ST-SynSet projection using standard distance metric such as Euclid distance. Therefore, we propose to learn a new distance metric for the visual words by integrating semantic constraints. We use the ST-SynSets structure to get the semantic constraints where we maintain that the visual words of the same ST-SynSet get closer in the new feature space than those in different ST-SynSets.

Given n visual words $\{x_1, \ldots x_n\}$, with all $x_i \in R^d$, we need to compute a positive-definite $d \times d$ matrix A to parameterize the squared Mahalanobis distance:

$$d_A(x_i, x_j) = (x_i - x_j)^T A(x_i - x_j) \qquad i, j = 1.....n \tag{5}$$

Till now, many methods have been proposed for Mahalanobis metric learning [15, 16, 17], and we utilize the information-theoretic metric learning method in [17] because it is fast and effective for the similarity constraints. Given an initial $d \times d$ matrix A_0 specifying the standard metric about inter-point distances, the learning task is posed as an optimization problem that minimizes the LogDet divergence between matrices A and A_0, subject to a set of constraints specifying pairs of examples that are similar or dissimilar.

Here we define the rule of the constraints to preserve small distances for visual words in the same ST-SynSet and large distances for those in different ST-SynSets. The problem is formalized as follows:

$$\begin{aligned} &\min_{A \geq 0} D_{ld}(A, A_0) = trace(AA_0^{-1}) - \log\det(AA_0^{-1}) - d \\ &s.t. \quad d_A(x_i, x_j) \leq u \quad (i, j) \in \textit{same ST-SynSet} \\ &\qquad d_A(x_i, x_j) \geq l \quad (i, j) \in \textit{different ST-SynSets} \end{aligned} \tag{6}$$

A_0 is unit matrix for a standard squared Euclidean distance, and l and u are respectively large and small values, which are given empirically according to the sampled distances of visual words. This problem can be optimized using an iterative optimization procedure by projecting the current solution onto a single constraint per iteration.

With the learned metric matrix A, the distance between local feature and visual word is computed as Equation (5). Then, we use the K-nearest neighbor algorithm to do ST-SynSet projection. The mapping to the ST-SynSet is measured by the majority vote of the feature's K nearest visual words. Here, K is empirically set to be 5.

2.3 Bag of ST-SynSets Action Classification Using SVM

Once the ST-SynSets and the new metric for visual words are obtained, we can describe a given action video using the histogram of ST-SynSets, in the way of "Bag of

ST-SynSets". We use SVM classifier to model each action category. Here, histogram intersection kernel in Equation (7) is used as the kernel.

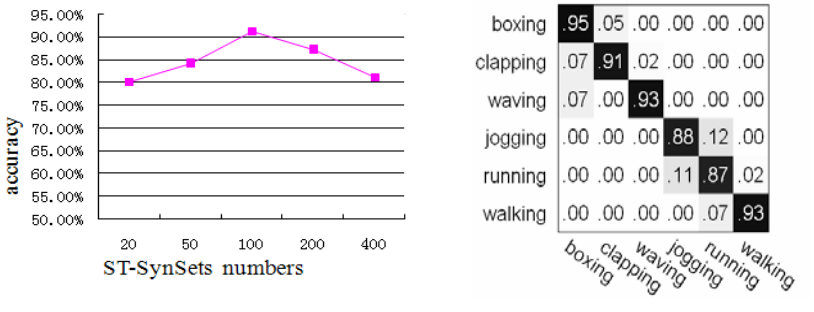
$$k_{HI}(h_1, h_2) = \sum_{i=1}^{n} \min(h_1(i), h_2(i)) \tag{7}$$

where h_1, h_2 are the histogram of two videos, and $h(i)$ is the frequency of the i^{th} bin.

3 Experiments

We test the proposed method on the KTH human motion dataset [7]. This dataset contains 598 short videos of six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed by 25 persons. The dataset is challenging because there are four different scenarios with variable backgrounds, moving camera and changed scales.

We extract spatio-temporal interest points using the detector proposed by Dollar et al. in [5] and get the corresponding gradient-based descriptors. The detector's scale parameters are empirically set with $\sigma = 2$ and $\tau = 3$ and PCA is applied to get a lower dimension of 100. We build visual vocabulary with the videos of randomly selected 3 persons for each action, and the initial number of visual words is set to 1000. Then we construct ST-SynSets using the SIB algorithm in Section 2.1.1, with a variable number of {20, 50, 100,200, 400} clusters, and after ST-SynSets selection we retain 80 percent of ST-SynSets with higher scores. We adopt the Leave One Out Cross Validation (LOOCV) as [6]. More specifically, at each run of the LOOCV, we use videos of 24 persons to train SVM and the rest for testing, and the average accuracy of 25 runs is reported as the results. The results of the algorithm with different ST-SynSets numbers are shown in Fig 5(a). We can see the optimal accuracy is obtained when the number of ST-SynSets is set to be 100, and after informative ST-SynSets refinement, the total number of ST-SynSets used to represent actions is 80.



(a) (b)

Fig. 5. (a) Average accuracy of the algorithm with different ST-SynSets numbers; (b) Confusion matrix of the method with the best ST-SynSets which are set to be 100 clusters and have an effective ST-Synsets of 80

Confusion matrix of the proposed method with the best ST-SynSets number is shown in Fig 5 (b). We can see most actions are recognized correctly, and the largest confusion is “jogging” vs. “running”, which share similar local features and are easily confused.

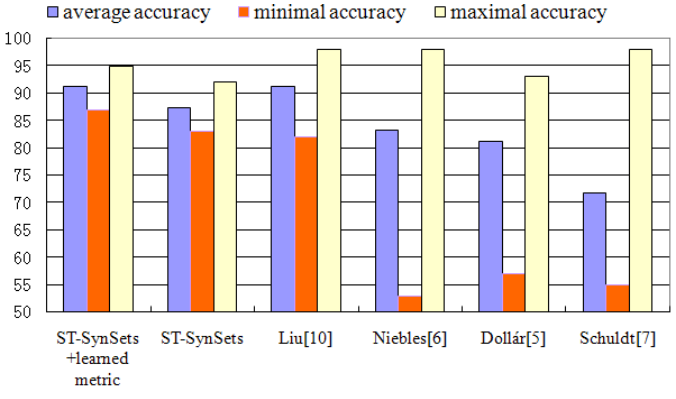


Fig. 6. Performance comparison with other methods

Fig. 6 shows the comparison of accuracies of the popular methods in recent years based on spatio-temporal local features. Although almost all the methods achieve high maximal accuracies for the “walking” action which has slight geometric and viewpoint changes in different scenarios, the proposed method shows superior performance in the minimal accuracy especially for the easily confused actions such as “running”. From Fig 6, we can see the proposed method with the ST-SynSet representation and the learned metric achieves the relatively most reliable and stable performance for all types of actions, with the average accuracy of 91.16% and the standard deviation of only 3.13%. The main reason is that this approach fully considers the category semantic information and gets the most discriminative representation visual words clusters to handle the intra-class local variations caused by different viewpoints as well as geometric and anthropometry variances. By clustering the visual words to a semantic meaningful unit with similar discriminative power and by learning the new metric for visual words using synonym constraints, it can learn a category semantic consistent feature space for the visual vocabulary and thus partially bridge the semantic gap in terms of significant intra-class variations and inter-class confusion.

In Fig.7, we show example videos in the four different scenarios from four confusable action categories with their corresponding ST-SynSets histograms when ST-SynSets number is set to be 20. We can see that actions from the same category share the similar ST-SynSets distributions. It is also clear to see from the peaks of these histograms that some ST-SynSets are dominating in particular actions and have the discriminative power among different actions.

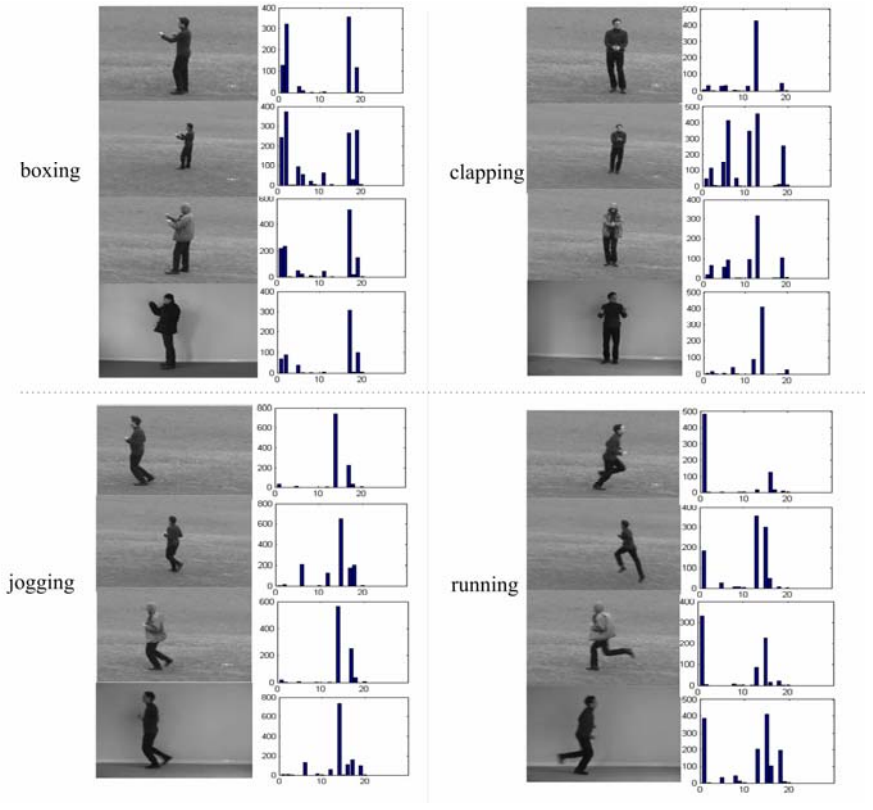


Fig. 7. Examples of ST-SynSets histograms for 4 confusable actions performed in the different scenarios

4 Conclusion and Future Work

In this paper, we propose a higher level representation for action recognition, namely Bag of Spatio-temporal Synonym Sets to bridge the gap between visual appearances and category semantics of human actions. By grouping the semantic-consistent visual words to be ST-SynSet and learning a new distance metric for visual words with both visual and semantic constraints, this approach can partially handle the intra-class variations and inter-class similarities. Experimental results on the KTH dataset show the effectiveness of the proposed method for human action recognition in the scenarios of different viewpoints and illumination conditions.

Because of the good results shown by other methods using the same constrained action database, further comparisons in some more challenging realistic databases with huge variations of anthropometry ,viewpoint, illumination, occlusions and so on are required to highlight the advantage of our method, namely, the robustness and discriminative power to handle the significant intra-class variations. In addition, since the ST-Synset is more compact than visual word, it has relatively stronger extension

ability for integrating global geometry and context information. All these will be addressed as our future work.

Acknowledgments. This work was supported by the National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60902090, 60802067), Beijing New Star Project on Science & Technology (2007B071), Co-building Program of Beijing Municipal Education Commission.

References

1. Bobick, A.F., Davis, J.W.: The Recognition of Human Movement Using Temporal Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 257–267 (2001)
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: IEEE International Conference on Computer Vision, vol. 2, pp. 1395–1402 (2005)
3. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision, vol. 2, pp. 726–733 (2003)
4. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: IEEE conference on Computer Vision and Pattern Recognition, pp. 405–412 (2005)
5. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, pp. 65–72 (2005)
6. Niebles, J.C., Wang, H.C., Li, F.F.: Unsupervised Learning of Human Action Categories using Spatial-Temporal Words. International Journal of Computer Vision 79, 299–318 (2008)
7. Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: A Local SVM Approach. In: International Conference on Pattern Recognition, vol. 3, pp. 32–36 (2004)
8. Savarese, S., DelPozo,Niebles, J.C., Li, F.F.: Spatial-temporal correlations for unsupervised action classification. In: IEEE Workshop on Motion and Video Computing (2008)
9. Boiman, O., Irani, M.: Detecting irregularities in images and in video. International Journal of Computer Vision 74(1), 7–31 (2007)
10. Liu, J., Shah, M.: Learning Human Actions via Information Maximization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
11. Winn, J., Criminisi, A., Minka, T.: Object Categorization by Learned Universal Visual Dictionary. In: IEEE International Conference on Computer Vision, pp. 1800–1807 (2005)
12. Zheng, Y.T., Zhao, M., Neo, S.Y., Chua, T.S.: Visual synset: towards a higher-level visual representation. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
13. Slonim, N., Friedman, N., Tishby, N.: Unsupervised document classification using sequential information maximization. In: Proceedings of the 25th ACM SIGIR international conference on research and development in information retrieval, pp. 129–136 (2002)
14. Lin, J.: Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory 37(1), 145–151 (1991)
15. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side-information. In: Advances in neural information processing systems, vol. 16, pp. 521–528 (2002)
16. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning a Mahalanobis Metric from Equivalence Constraints. Journal of Machine Learning Research 6, 937–965 (2005)
17. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-Theoretic Metric Learning. In: International Conference on Machine Learning, pp. 209–216 (2007)