Recognizing realistic action using contextual feature group

Yituo Ye[1], Lei Qin[2], Zhongwei Cheng[1], Qingming Huang[1, 2, *]

[1]Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
[2] Key lab of Intelli. Info. Process., Inst. of Comput. Tech., CAS, Beijing 100190, China
E-mail: {ytye, lqin, zwcheng, qmhuang}@jdl.ac.cn

Abstract. Although the spatial-temporal local features and the bag of visual words model (BoW) have achieved a great success and a wide adoption in action classification, there still remain some problems. First, the local features extracted are not stable enough, which may be aroused by the background action or camera shake. Second, using local features alone ignores the spatial-temporal relationships of these features, which may decrease the classification accuracy. Finally, the distance mainly used in the clustering algorithm of the BoW model did not take the semantic context into consideration. Based on these problems, we proposed a systematic framework for recognizing realistic actions, with considering the spatial-temporal relationship between the pruned local features and utilizing a new discriminate group distance to incorporate the semantic context information. The Support Vector Machine (SVM) with multiple kernels is employed to make use of both the local feature and feature group information. The proposed method is evaluated on KTH dataset and a relatively realistic dataset YouTube. Experimental results validate our approach and the recognition performance is promising.

Keywords: Action recognition, Spatial-temporal local feature, Local feature group, Discriminate group distance, Mahalanobis distance.

1 Introduction

Recognizing human activities is one of the most promising fields of computer vision. It is receiving increasing attention due to its wide range application such as smart surveillance, human-computer interface, virtual reality, content based video retrieval and video compression.

Although a large amount of research has been reported on human actions recognition, there still remain some open issues, and one of the most important issues is action representation. Among the traditional approaches, holistic information is always used to model human actions. Bobick and Davis [1] proposed the MEI and MHI method, which is capable of encoding the dynamics of a sequence of moving human silhouettes. Ke *et al.* [2] used segmented spatial-temporal volumes to model human activities. Although their

* Corresponding author.

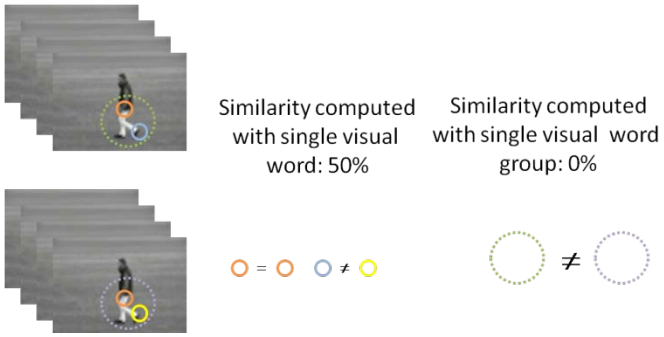


Fig. 1. The quantization error are magnified when combining visual words together, resulting in the accuracy reduction

methods are efficient, most of the holistic approaches have either the requirement of pre-processing or an expensive computational cost. Due to the limitation of holistic approaches, part-based approaches which only use several 'interesting' parts received more attention. And one of the most popular presentations is the Bag of Visual Words model (BoW) [3, 5]. The procedure of BoW is clustering a large number of local features to make visual words vocabulary and then quantizing different features to their corresponding nearest visual words.

Notwithstanding its great success and wide adoption in BoW, this method still has some issues. First, single visual word discards rich spatial information among local features, which is of great importance in human activities recognition. Some previous works [8. 9, 10] have verified that modeling these visual contexts can improve the performance. The common approach is trying to identify the combination of visual words with statistically stable spatial configurations. Liu *et al.* [8] utilized feature pursuit algorithms such as AdaBoosting to model the relationship of visual words. Ryoo and Aggarwal [9] introduced the spatial-temporal relationship match (STR match), which considers spatial and temporal relationship among detected features to recognize activities. Hu *et al.* [10] proposed the spatial-temporal descriptive video-phrases (ST-DVPs) and descriptive video-clips (ST-DVCs) to model the spatial and temporal information of visual words. Generally, model the relationship of visual words can benefit the recognition. However, the quantization error introduced during visual vocabulary generation may degrade the accuracy, which can be seen in figure 1. The method proposed in this paper models the spatial and temporal information of local features rather than the visual words to avoid the influence of the quantization error.

Second, the distance metric, such as Euclidean distance and L1-norm, commonly used for generating visual vocabulary, does not take the semantic context into consideration. This may render them to prone to noise, for that the local features with similar semantic could be clustered in different visual words. Inspired by the metric learning framework of Zhang *et al.* [11], we present a new spatial context weighted Mahalanobis distance metric to measure the similarity between different features, and furthermore, the group distance

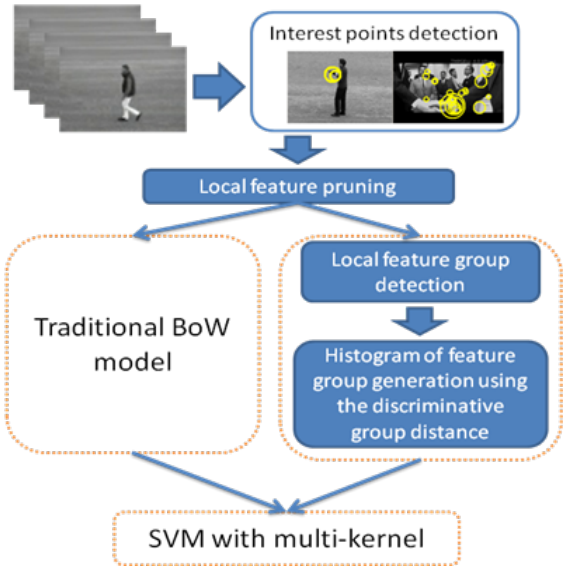


Fig. 2. The proposed framework

can also be computed based on it, which is named as discriminate group distance.

Based on these problems, we propose an action recognition framework, which utilizes the group of local features to construct visual vocabulary, and then quantizes the new group to its corresponding nearest word, using the proposed new discriminate group distance. Combining with the histogram of single features, different actions can be recognized using the SVM classifier. The whole process is illustrated in figure 2.

The remainder of this paper is organized as follows. Section 2 introduces our proposed local feature group detector. Section 3 illustrates the discriminate group distance and the corresponding classifier. Section 4 presents and discusses our experimental results on KTH and YouTube dataset. We conclude the paper in Section 5.

2 Local feature group detection and representation

To extract the local feature group, we first utilize the 3D-Harris [3] detector to detect the interest points, and use the HOF and HOG as in [3] to represent interest points. For each interest point, two kinds of information can be acquired, the local feature descriptor D and the scale information S. Then each local feature can be denoted as $F(S, D)$, and a local feature group can be represented as $G\{F_1, F_2, F_3 \ldots F_n\}$, where n is the number of features in a group.

To make local feature group representative and robust, group extraction algorithm should be accord with some rules: 1) feature group should be robust to noise, such as background action and camera shake; 2) feature group should be scale invariant; 3) the

Fig. 3. The feature group detector. It extracts the local features positioned in the radius *R*, and constructs the feature groups.

number of local features in a feature group should be small. Feature pruning algorithm proposed in [8] is adopted to make the extracted features satisfy the rule 1. The group extraction method will be discussed in section 2.1, which has taken into account of the rule 2 and rule 3.

2.1 Local feature group extraction

Different algorithms have been proposed to detect local feature groups [8, 9, 10]. In this paper, we define the co-occurred local features, which satisfy certain spatial-temporal restriction, as a feature group. In order to satisfy the second rule shown above, we use the scale information as the basis to compute the spatial-temporal distance between local features. As for the third rule, if too many local features are combined, the repeatability of the local feature group will decrease. Furthermore, if more local features are contained in a local feature group, there would be more possible feature-to-feature matches between two groups, which would make the computation of group distance time consuming. As a tradeoff, we fix the number of local features in each local feature group as 2 in this paper.

To detect local feature group, we use the detector illustrated in figure 3. In this figure, a sphere with radius R is centered at a local feature. A local feature group is formed by the centered local feature and other local features within the sphere. To make the feature group invariant to scale, the radius is set as

$$R = S_{center} \times \lambda \tag{1}$$

where S_{center} is the scale of centered local feature and λ is a parameter that controls the spatial-temporal span of local feature group. A large λ will overcome the sparseness of local feature group and indentify stable spatial-temporal relationship between local features. However, a larger λ also requires more computational cost.

By scanning every local feature with the detector, the local feature groups, each of which contains two local features, are generated. It should be noted that, the new group contains different features rather than different quantized visual words, which makes it robust to quantization error.

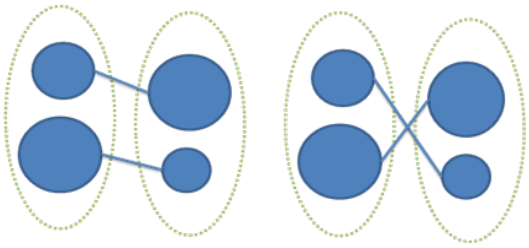


Fig. 4. The possible match orders when each local group contains two local features

By extracting local features and feature groups, two kinds of information can be acquired. They are further passed to the BoW model with a new discriminate group distance. Then we can get two histograms, corresponding to the local features and feature groups. Finally, different actions can be classified by SVM with multiple kernels.

Section 3.1 will introduce the discriminate group distance, which is a combination of context weighted Mahalanobis distance metric. Section 3.2 will present the related SVM classifier. With two kinds of information, multi-kernel learning strategy is employed to improve the recognition accuracy.

3.1 Discriminate group distance and metric learning

The discriminate group distance is defined as a combination of spatial-temporal context weighted Mahalanobis distance between two feature groups. Note that discriminate group distance is computed between groups containing identical number of local features, so there are $n!$ feature to feature matches when each group contains n local features. As illustrated in figure 4, when $n = 2$, there are two possible matches.

In [11], a best match order is defined as the one that maximize the spatial similarity based on the scale and orientation information. As for this paper, we take every match into consideration and select the one with the minimal distance. And as in figure 4, the second match order should be chosen. When $n = 2$, the discriminate group distance can be represented as

$$GD(G_I, G_J) = \min\left(\sum_{k=1}^{2} d(D_I^k, D_J^k), \sum_{k=1}^{2} d(D_I^k, D_J^{3-k})\right) \tag{2}$$

In the formula above, $GD(G_I, G_J)$ denotes the discriminate group distance between G_I and G_J , D^k_I is the kth local feature in group I and $d(D^k_I, D^k_J)$ represents the Mahalanobis distance between feature D^k_I and D^k_J.

As for the $d(D^k_I, D^k_J)$ shown above, a Mahalanobis distance is utilized which could incorporate the semantic context between local feature groups. Thus, it can be represented as

$$d(D_I^k, D_J^k) = (D_I^k - D_J^k)^T A (D_I^k - D_J^k) \tag{3}$$

For we use HOF and HOG to present the local feature as [3], A is a 144×144 matrix to be learned from the semantic labels of the local feature groups.

Intuitively, we try to find a good distance metric which makes the feature groups with similar semantic contexts close to each other and those with different semantic appearing far away. To achieve this, metric learning algorithm proposed in [11] is applied, whose result is acquired through iterative calculation when given a set of local features and their corresponding labels.

3.2 Multi-kernel SVM classifier

For classification, we use a non-linear SVM with the histogram intersection kernel, which can be presented as

$$K(H_I, H_J) = \sum_{k=1}^{N} \min(h_I^k, h_J^k) \tag{4}$$

where $H_I = \{h_I^n\}$ and $H_J = \{h_J^n\}$ are the histograms either for the local features or the feature groups.

Note that, there are two kinds of histograms. To combine the information effectively, two methods are utilized: *multi-kernel1* (MK_1) and *multi-kernel2* (MK_2), which can be represented respectively as

$$MK_1 = \frac{(K_1(H_I, H_J) + K_2(H_I, H_J))}{2} \tag{5}$$

$$MK_2 = \sqrt{K_1(H_I, H_J) \times K_2(H_I, H_J)} \tag{6}$$

$K_1(H_I, H_J)$ and $K_2(H_I, H_J)$ are the kernel matrices for local features and feature groups respectively.

4 Experiments

We have tested the proposed methods on the dataset KTH and YouTube, and the results prove that the methods can enhance the action recognition performance. Section 4.1 will briefly introduce the dataset KTH and YouTube, Section 4.2 and Section 4.3 would present and discuss the experiment on KTH and YouTube respectively.

4.1 KTH and YouTube dataset

KTH is a relatively simple dataset which contains about 600 videos performed by 25 actors. ‘walking’, ‘jogging’, ‘running’, ‘boxing’, ‘hand waving’ and ‘hand clapping’ are the six actions in the dataset. Videos are taken at slightly different scales with various

Fig. 5. Examples of the YouTube dataset

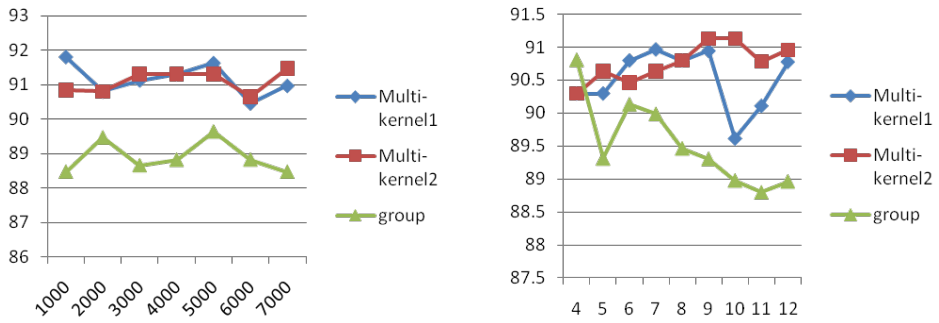
backgrounds, indoor and outdoor environments. Each video contains repeated executions of a single action in a resolution of 160×120, 25fps.

YouTube is a dataset "in the wild", its source is YouTube videos and the videos are collected by Liu *et al.* [8]. It contains about 1160 videos and includes 11 categories: 'basketball shooting', 'volleyball spiking', 'trampoline jumping', 'soccer juggling', 'horseback riding', 'cycling', 'diving', 'swing', 'golf swinging', 'tennis swinging' and 'walking with a dog'. Figure 5 gives a brief impression, from which we can drive some visualized properties such as the cluttered background, variations in object scale, varied viewpoints and varied illuminations. Besides, the videos also mix steady cameras and shaky cameras, which make the noise pruning even more necessary.

4.2 Experiments on KTH dataset

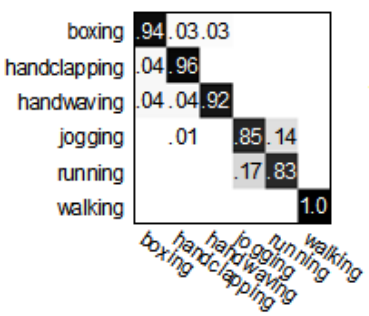
Since the KTH datasets is relatively "clean", feature pruning is not necessary. We performed two groups of experiments. The first one is to test the selection of the parameter λ in feature group extraction, and the second one is to test the selection of the parameter K in the clustering algorithm k-means of vocabulary construction. As for the training set and testing set, we apply the leave-one-out-cross-validation (LOOCV) scheme and use the mean value as the accuracy.

Traversing every possible value of the pair (λ, K) in the procedure will be time consuming. So in our experiment, the selection of λ will be tested with a constant value of K, and the selection of K will be tested with a stable λ. The results are illustrated on figure 6, from which *group* stands for the accuracy using only local feature group information, and *multi-kernel1* and *mutli-kernel2* represent the accuracy acquired by the mutli-kernel SVM classifier as illustrated in Section 3.2 It can be seen that combining



(a) (b)

Fig. 6. Influence of parameters, (a) denotes the relationship between average accuracy and parameter K when $\lambda = 8$, (b) implies the influence of parameter λ, when K= 2000.



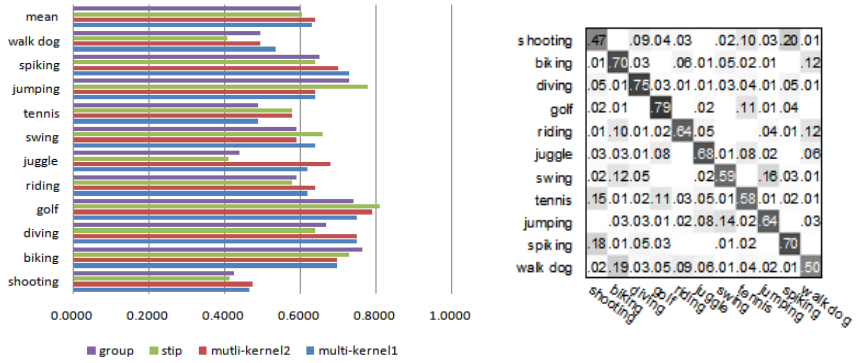
Method	Accuracy
J. Dollar et al.[4]	81.17%
J. Niebles et al[14].	83.3%
G. Taylor[15]	90.0%
S. Ji [13]	90.2%
Y Wang et al. [12]	91.20%
Ours method	91.60%

Fig. 7. Confusion table for the method we proposed and the comparison with other literature about the result on KTH using LOOCV scheme

local feature and feature group information will get a better accuracy. Meanwhile, with the growth of λ, the tendency of the accuracy is rising while computational time is also increasing. On the KTH dataset, 9 and 5000 are assigned to λ and K respectively and the performance of the proposed method is comparable with the state-of-the-art result on KTH which is 91.2% [12].

4.3 Experiments on YouTube dataset

For the YouTube dataset, the recognition is also carried out through the LOOCV scheme. Feature points of YouTube videos are processed by the pruning method, which leads to a nearly 2% improvement of recognition accuracy, *i.e.* from 58.1% to 60.5%. We use the parameter λ = 4 and K = 2000 in these experiments.



(a) (b)

Fig. 8. Results on YouTube dataset. (a)The average accuracy for multi-kernel1, multi-kernel2, BoW and group methods are 63.07%, 63.97%, 60.47% and 59.87% respectively, and (b) denotes the confusion table of multi-kernel2.

Then we verified the effectiveness of the method we proposed. MK1 reaches the accuracy about 63.07% , MK2 performs best and reaches nearly 63.97%, which is comparable of the state-of-the-art 65.4% [8] and superior to the 60.47% of traditional BoW model, using group information alone acquires the accuracy 59.87%. Based on figure 8(a), it can be seen that the traditional BoW model gets poor results on the action 'walk dog', 'juggle' and 'shooting', but there would be an improvement varied between 6% to 27% for MK2. It may be aroused by the feature group information which considers the spatial and temporal information between the local features extracted from 'human', 'basketball', 'football' and 'dog'. Taking 'walking dog' for example, the position of 'dog' and 'human' are always obey some spatial restriction, while the traditional BoW model which only presents the distribution of the local features discards this spatial information. However, it also should be noted that the feature group information may bring some kinds of spatial temporal restriction and may decrease the accuracy on some action with large intra-class variation such as 'swing'. On the whole, features group would be the complement for the local features rather than a replacement, combing these two kinds of information may acquire promising result. Figure 8 (b) shows the confusion matrix, we can see that 'biking' and 'riding horse' which share the similar motion are often misclassified as 'walking dog', 'shooting' and 'volleyball spiking' are often confused due to their common action jumping with a ball.

5 Conclusions

We propose a systematic framework for recognizing realistic actions, which considers the spatial-temporal relationship between the local features and utilizes a new discriminate group distance using a combination of the Mahalanobis distance for the clustering

algorithm in the BoW model. The effectiveness has been tested on the datasets KTH and YouTube. Experimental results verify that our framework is effective, and combing the spatial temporal information between local features improves the recognition accuracy.

Acknowledgements

This work was supported in part by National Basic Research Program of China (973 Program): 2009CB320906, in part by National Natural Science Foundation of China: 61025011, 61035001 and 61003165, and in part by Beijing Natural Science Foundation: 4111003.

References

1. A. Bobick and J. Davis. The representation and recognition of action using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3):257–267, 2001.
2. Ke Y, Sukthankar R and Hebert. Spatial-temporal shape and flow correlation for action recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2007.
3. I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.
4. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.
5. L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.
6. P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In ACM International Conference on Multimedia, 2007.
7. H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In IEEE International Conference on Computer Vision, 2007.
8. Jinen Liu, Jiebo luo, Mubarak Shah. Rcognizing realisitic actions from videos “in the wild”. In IEEE Conference on Computer Vision and Pattern Recognition, 2009
9. RYoo, M.S and AGGARWAL, J.K. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In IEEE International Conference on Computer Vision, 2009.
10. Qiong Hu, Lei Qin, Qingming Huang, Shuqiang Jiang, and Qi Tian. Action Recognition Using Spatial-Temporal Context," Analysis and Search. 20th International Conference on Pattern Recognition, August 23-26, 2010, Istanbul, Turkey.
11. Shiliang Zhang, Qingming Huang, Gang Hua, Shuqiang Jiang, Wen Gao, and Qi Tian, Building Contextual Visual Vocabulary for Large-scale Image Applications, in Proceedings of ACM Multimedia Conference, *ACM MM* (Full Paper), Florence, Italy, pp.501-510, Oct.25-29, 2010.
12. Y Wang, G Mori, “Human Action Recognition by Semi-Latent Topic Models,” PAMI, 2009.
13. S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. In *ICML*, 2010. 3362, 3366
14. J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categpries using sptatial-temporal words. IJCV, 2008. 3366.
15. G. Taylor, R. Fergus, Y. Lecun, and C. Bregler. Convolutional learning of spatio-temporal features. In ECCV, 2010. 3361, 3362, 3366, 3367.