IEICE Electronics Express, Vol.7, No.1, 7–12

Video annotation based on adaptive annular spatial partition scheme

Guiguang Ding[a)], **Lu Zhang, and Xiaoxu Li**

Key Laboratory for Information System Security, Ministry of Education, Tsinghua National Laboratory for Information Science and Technology, School of Software, Tsinghua University, Beijing 100084, China

a) *dinggg@tsinghua.edu.cn*

Abstract: The method based on Bag-of-visual-Words (BoW) deriving from local keypoints has recently appeared promising for video annotation. Spatial partition scheme has critical impact to the performance of BoW method. In this paper, we propose a new adaptive annular spatial partition scheme. The proposed scheme firstly determines the centroid of partition according to the distribution of keypoints. And then the image is partitioned into several annular regions. In the end, BoW histograms are computed according to the annular regions, which are utilized to train SVM classifiers. A systematic performance study on TRECVID 2006 corpus containing 20 semantic concepts shows that the proposed scheme is more effective than other popular spatial layout partition schemes such as 2×2 grid scheme.

Keywords: video annotation, Bag-of-visual-Words, spatial partition

Classification: Science and engineering for electronics

References

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Y. G. Jiang, C. W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," *ACM International Conference on Image and Video Retrieval (CIVR'07)*, Amsterdam, Netherlands, 9-11, July 2007.

[3] Y. Liang and X. Liu, et al., "THU and ICRC at TRECVID 2008," in *TRECVID workshop*, 2008.

[4] S.-F. Chang and J. He, et al., "Columbia University/VIREO-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search," in *TRECVID workshop*, 2008.

[5] TREC Video Retrieval Evaluation (TRECVID). [Online] http://www-nlpir.nist.gov/projects/trecvid/

DOI: 10.1587/elex.7.7
Received May 13, 2009
Accepted December 06, 2009
Published January 10, 2010

1 Introduction

The rapid growth of the video data has created a compelling need for innovative tools to retrieve and manage the large video collections. One major challenge in video retrieval and management is to bridge the so-called "semantic gap" between low-level features and high-level semantic concepts. The semantic classifier based on Bag-of-visual-Words (BoW) is one of the most effective methods to tackle the semantic gap, which has recently attracted numerous research attentions. The basic idea of BoW is to depict each image as an orderless collection of local keypoint features. To represent images with the BoW, three main things need to be considered: extract the local features, construct visual vocabulary and map local features in an image to the visual vocabulary. Fig. 1 shows the framework of the video annotation method based on BoW. The local features are firstly extracted from salient image patches by keypoint detector and descriptor (e.g. DoG + SIFT [1]). Then a visual vocabulary is constructed through a clustering algorithm (e.g. K-means) to cluster the local keypoints. Each keypoint cluster is treated as a visual word in the visual vocabulary. By mapping the keypoints in an image to the visual vocabulary, an image can be represented as a feature vector with each dimension corresponding to a visual word. The BoW representation has appeared promising for semantic concept detection of video [2].

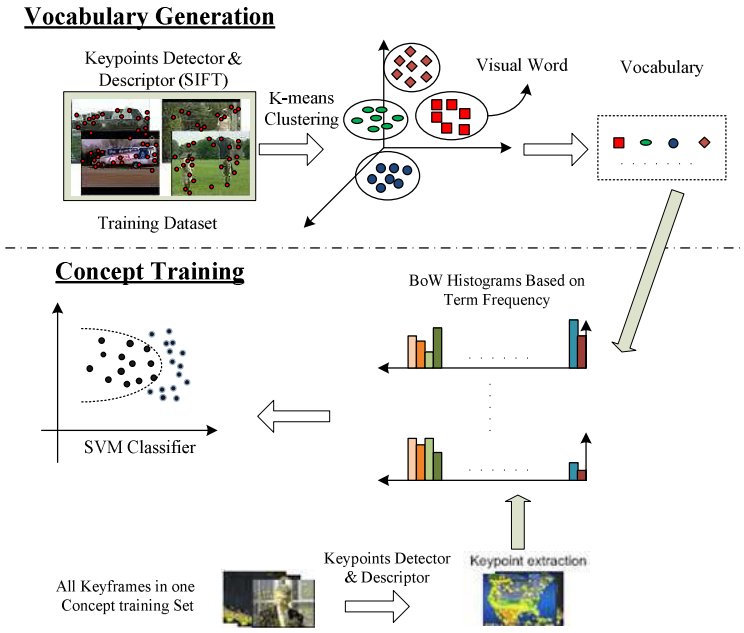

Fig. 1. Framework of the video annotation method based on BoW.

The spatial locations of keypoints in an image carry important information for video annotation. The spatial layout partition scheme is one of the most important impact to the performance of BoW. The scheme is based on the following assumptions: (1) videos are of similar capturing customs and styles, (2) most of the view points are with the horizontal line, and (3) most

DOI: 10.1587/elex.7.7
Received May 13, 2009
Accepted December 06, 2009
Published January 10, 2010

of the target objects are captured in the center of the screen [3]. In [4], several equal-size rectangular partition schemes: 1×1 (whole frame), 2×2, 3×3, and 4×4 were used. Their experiments in TRECVID-2006 test data showed that relatively coarse partitions of image regions (such as 2×2) were more effective than fine-grained partitions (such as 3×3 or 4×4). This is mainly due to the fact that their rectangular partition scheme could make many objects cross region boundaries and cause feature mismatch problem [4]. To overcome the drawbacks of the equal-size rectangular partition scheme, in this paper, we propose an adaptive annular spatial partition scheme. Experiments demonstrate the proposed scheme outperforms significantly equal-size rectangular partition schemes in video annotation applications.

2 The Proposed scheme

In this section, we describe the proposed spatial partition scheme based on the distribution of keypoints. Like other BoW methods, given a set of keypoints, we first construct a visual vocabulary through clustering the keypoints by k-means algorithm, as shown in Fig. 1. Then we divide the keypoints in an image into three annular regions, compute the visual-word feature (BoW histogram) for each region, and concatenate the features of these regions into an overall feature vector. The overall feature vector is used as the training data of SVM classifier to acquire the classifier for each semantic concept. Fig. 2 presents the framework of the adaptive annular spatial partition scheme.

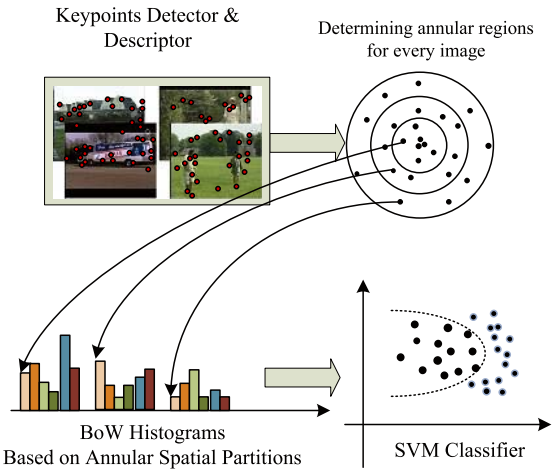

Fig. 2. Framework of the proposed annular spatial partition scheme.

The proposed scheme adaptively determines the centroid and the annular regions according to the distribution of keypoints. The processing is described as follows:

Let $P = \{p(x, y)\,|1 \le x \le R; 1 \le y \le C\}$ be the set of the interest points in an image of size $C \times R$. Let N_p is the number of the interest points. Let

DOI: 10.1587/elex.7.7
Received May 13, 2009
Accepted December 06, 2009
Published January 10, 2010

$M = (\overline{x}, \overline{y})$ be the centroid of P, where $\overline{x}$ and $\overline{y}$ are defined as:

$$\overline{x} = \frac{1}{N_p} \sum_{(x,y)\in P} x; \quad \overline{y} = \frac{1}{N_p} \sum_{(x,y)\in P} y \tag{1}$$

Let r be the radius of P which is defined as:

$$r = \max_{(x,y)\in P} \sqrt{(x-\overline{x})^2 + (y-\overline{y})^2} \tag{2}$$

Given $N = 3$, uniformly divide the radius into N buckets, then draw N concentric circles with M as the center and with $\frac{kr}{N}(k = 1, 2, 3)$ as the radius to form three annular regions. The set of interest points in every annular region is:

$$R_k = \left\{ (x,y) \left| \frac{(k-1)r}{N} \le \sqrt{(x-\overline{x})^2 + (y-\overline{y})^2} \le \frac{kr}{N} , (x,y) \in P \right. \right\}, \quad k = 1, 2, 3 \tag{3}$$

Fig. 3 shows the partition result of the annular regions for an airplane image.

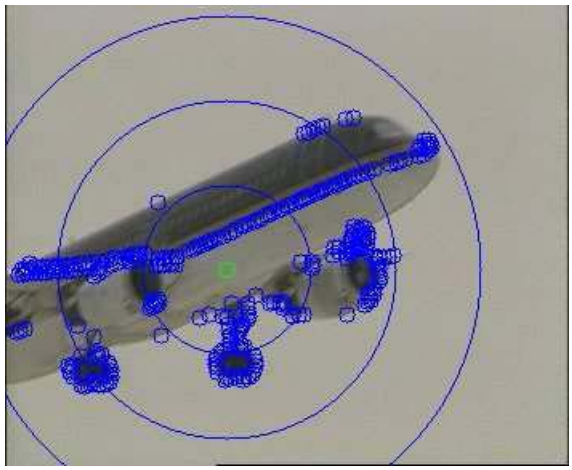

Fig. 3. Partition result of the annular regions.

After determining annular regions, we map the keypoints to the visual vocabulary and compute the BoW histogram for each region by considering the frequency of each visual word appearance in a region. In the end, we concatenate the BoW histograms from three regions into an overall BoW histogram according to the following equation.

$$H_{overall} = \alpha \times H_1 \mapsto \beta \times H_2 \mapsto \gamma \times H_3 \tag{4}$$

Where the symbol $\mapsto$ represents to concatenate two histogram, H_1, H_2 and H_3 represent the BoW histogram of each region and α, β and γ represent the weight corresponding to every histogram. Because most of the target objects are captured in the centric region, the contribution of centric region's histogram should be larger than that of other regions' histogram. Generally, we let $\alpha > \beta > \gamma$.

DOI: 10.1587/elex.7.7
Received May 13, 2009
Accepted December 06, 2009
Published January 10, 2010

3 Experimental Results

To evaluate the proposed scheme for video annotation, we conduct the experiments on the benchmark video corpus of the TRECVID 2006 dataset [5], which consists of 137 broadcast news videos. These training videos are segmented into 61901 video shots and 39 concepts are labeled on each shot. In the experiments, 20 concepts are selected, and the keypoints are detected by DoG detector and described by the PCA-SIFT descriptor. We use the k-means clustering algorithm to generate the visual word vocabulary containing 2000 visual words. For all the key frames, the BoW features are calculated respectively based on the 1×1 spatial partition scheme, the 2×2 spatial partition scheme, and the proposed adaptive annular partition scheme. For each semantic concept, three SVM classifiers are trained respectively using the three BoW features.

For performance evaluation, we use non-interpolated Average Precision (AP) as the performance metric, which is the official performance metric in TRECVID. It reflects the performance on multiple average precision values along a precision-recall curve. In the experiments, the weight parameters are empirically chose as $\alpha = 1$, $\beta = 0.7$, $\gamma = 0.5$. Fig. 4 illustrates the AP of the three schemes. For each setting, the linear and RBF kernel of SVM are used, and the performance of the better one is reported. We can see that the annular partition scheme outperforms other two schemes for over 11 of all 20 concepts. Some of the improvements are significant, such as the 50% improvements on "animal". The proposed scheme remains no change on a few concepts. The main reason is that the training images about these concepts have no obvious spatial characteristics. In terms of the mean average precision (MAP), the annular partition scheme (9.4%) achieves around 15% improvements compared to the 1×1 partition scheme (8.2%). The results demonstrate that it is effective that the proposed annular partition scheme is used to partition an image for video annotation.

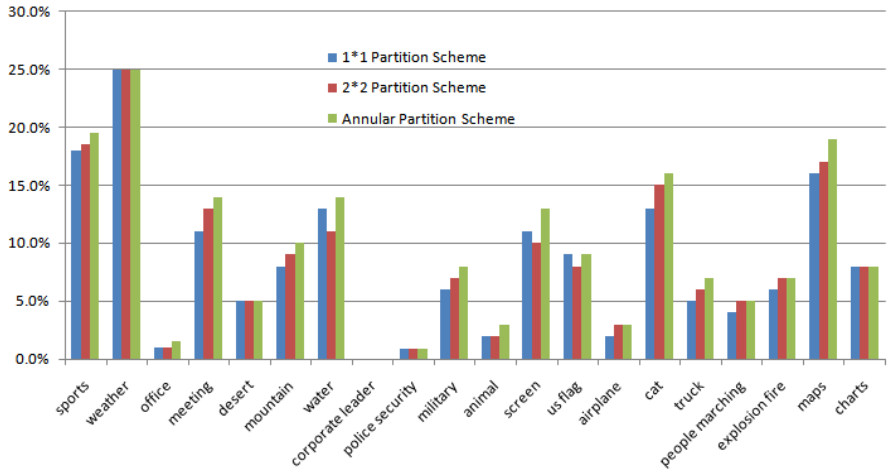

Fig. 4. Performances of different spatial partition schemes over all 20 concepts.

DOI: 10.1587/elex.7.7
Received May 13, 2009
Accepted December 06, 2009
Published January 10, 2010

4 Conclusion

In this paper, we have presented a novel spatial partition scheme for video annotation based on the distribution of keypoints. By partitioning an image into several annular regions, the effect of centric region is improved in concept classifiers. Experiments on the benchmark TRECVID data set demonstrated that the proposed scheme was superior to the conventional spatial partition scheme.

Acknowledgments

The authors acknowledge the support received from the National Natural Science Foundation of China (Project 60972096) and National 863 Plans Projects (Grant No. 200901Z410).

DOI: 10.1587/elex.7.7
Received May 13, 2009
Accepted December 06, 2009
Published January 10, 2010