



A visual word weighting scheme based on emerging itemsets for video annotation

Guiguang Ding*, Jianmin Wang, Kai Qin

School of Software, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:

Received 10 June 2009

Received in revised form 26 October 2009

Accepted 26 May 2010

Available online 1 June 2010

Communicated by A. Tarlecki

Keywords:

Information retrieval

Video annotation

Emerging itemsets

BoW

ABSTRACT

The method based on Bag-of-visual-Words (BoW) deriving from local keypoints has recently appeared promising for video annotation. Visual word weighting scheme has critical impact to the performance of BoW method. In this paper, we propose a new visual word weighting scheme which is referred as emerging patterns weighting (EP-weighting). The EP-weighting scheme can efficiently capture the co-occurrence relationships of visual words and improve the effectiveness of video annotation. The proposed scheme firstly finds emerging patterns (EPs) of visual keywords in training dataset. And then an adaptive weighting assignment is performed for each visual word according to EPs. The adjusted BoW features are used to train classifiers for video annotation. A systematic performance study on TRECVID corpus containing 20 semantic concepts shows that the proposed scheme is more effective than other popular existing weighting schemes.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

With the development of video devices and video processing technologies, the amount of video data has grown rapidly and enormously for various usages, such as advertising, news video broadcasting, personal video archive, medical video data, and so on [1]. 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 a feature vector related to local keypoints. To represent images with the BoW, there are three main steps: extract the local features, construct visual vocabulary and map local features

in an image to the visual vocabulary. The local features are firstly extracted from salient image patches by keypoint detector and descriptor (e.g. DoG + SIFT [2]). Then, by a clustering algorithm (e.g. K-means), the local keypoints are clustered to construct a visual vocabulary. In the visual vocabulary, each visual word is defined as a keypoint cluster. According to the visual vocabulary, an image can finally be represented as a feature vector whose each dimension corresponds to a visual word.

The BoW representation of image has appeared promising for semantic concept detection of video. The visual word weighting scheme is one of the most important impacts to the performance of BoW. The main weighting schemes with BoW are TF (Term Frequency) weighting, TF-IDF (TF-Inverse Document Frequency) weighting, binary weighting and soft-weighting [3–5]. These weighting schemes evaluate visual similarity by independently considering each visual word and bin-to-bin comparison of visual word histograms. They all neglect the co-occurrence relationships of visual words for one single concept, which is usually critical for the performance of video annotation. For example in Fig. 1, K_1 and K_2 are two positive examples for the concept “boat”, and K_3 is negative one. The weights of four visual words are given according to the

* Corresponding author.

E-mail addresses: dinggg@tsinghua.edu.cn (G. Ding), Jimwang@tsinghua.edu.cn (J. Wang), nicholas.qk@gmail.com (K. Qin).




Visual Words		Keyframes					
		...	W_1	W_2	W_3	W_4	...
K_1		...	6	4	0	2	...
K_2		...	4	7	4	2	...
K_3		...	5	7	6	0	...

Fig. 1. The BoW representation of keyframe using TF scheme.

TF weighting scheme. The similarity of K_2 and K_1 is lower than one of K_2 and K_3 using histogram bin comparison. However, for the concept “boat”, K_2 and K_1 are more alike in visual content. For this concept, W_1, W_2 and W_4 are an emerging pattern but this cue is not used in the TF weighting scheme. If keyframe includes one emerging pattern of the concept “boat”, it is obvious that the visual words belonging to this pattern should be more important than others for the concept “boat”. These visual words should be given larger weights. All existing schemes do not consider the co-occurrence relationships of visual words and cannot solve the problem illustrated in Fig. 1.

Emerging Patterns (EPs) are those itemsets whose supports in one class are significantly higher than their supports in the other class. A representative instance of the class should contain strong EPs of the same class. Motivated by the EPs mining approach which can be used to effectively capture the co-occurrence relationships of items from datasets, in this paper, we propose a novel EPs weighting (**EP-weighting**) scheme, which aims to utilize EPs to mine the co-occurrence relationships of visual words and to construct the effective BoW representation of keyframe. The main contributions of this paper can be summarized as follows: (1) To our knowledge, we are the first to utilize the EPs mining technique to the BoW representation of video/image; (2) we develop an effective weighting adjustment algorithm based on EPs to assign the importance of each visual word in BoW. The new scheme can capture the co-occurrence relationships of visual words to improve the performance of video annotation. The experimental results show that the new scheme outperforms alternative schemes in video annotation applications.

The rest of our paper is organized as follows. Section 2 lists some related work. Section 3 describes the proposed approach, and illustrates the key intuitions and main algorithms involved. Some simulation results are presented

in Section 4 and the conclusion of this paper is made in Section 5.

2. Related work

Automatic semantic annotation (sometimes it is also named concept detection or high-level feature extraction) of video is an efficient method to retrieve and manage the large video collections. Recently, the BoW representation deriving from local keypoints has shown remarkable performance for video annotation. Keypoints are salient patches that contain rich local information. Keypoints can be detected using various detectors and depicted by various descriptors. The keypoint detectors and descriptors are surveyed in [6] and [7] respectively. There have been many works using BoW representation for video annotation [9, 10, 15]. In TREC Video Retrieval Evaluation (TRECVID) 2008 [8], most of the methods submitted are based on the BoW representation [8–10], which exhibits surprisingly good performance for video annotation. In [3, 5], Y.G. Jiang et al. integrated and evaluated several factors which could impact the performance of BoW. These factors include the choices of keypoint detector, size of visual vocabulary, weighting scheme of visual words, and kernel function used in supervised learning. Our work studies the importance of weighting scheme of visual words and develops novel visual word weighting scheme for video annotation application.

EPs [11] are defined as multivariate features whose supports change significantly from one class to another. Since EPs capture the knowledge of sharp differences between data classes, effective classification systems based on EPs have been built [12, 13], which are competitive to other existing state-of-the-art classifiers. In [13], H. Fan et al. use EPs to construct the reliable weighting SVM model, which focuses on finding the true noise distribution in the training data to reduce the affects of noise and outliers. In

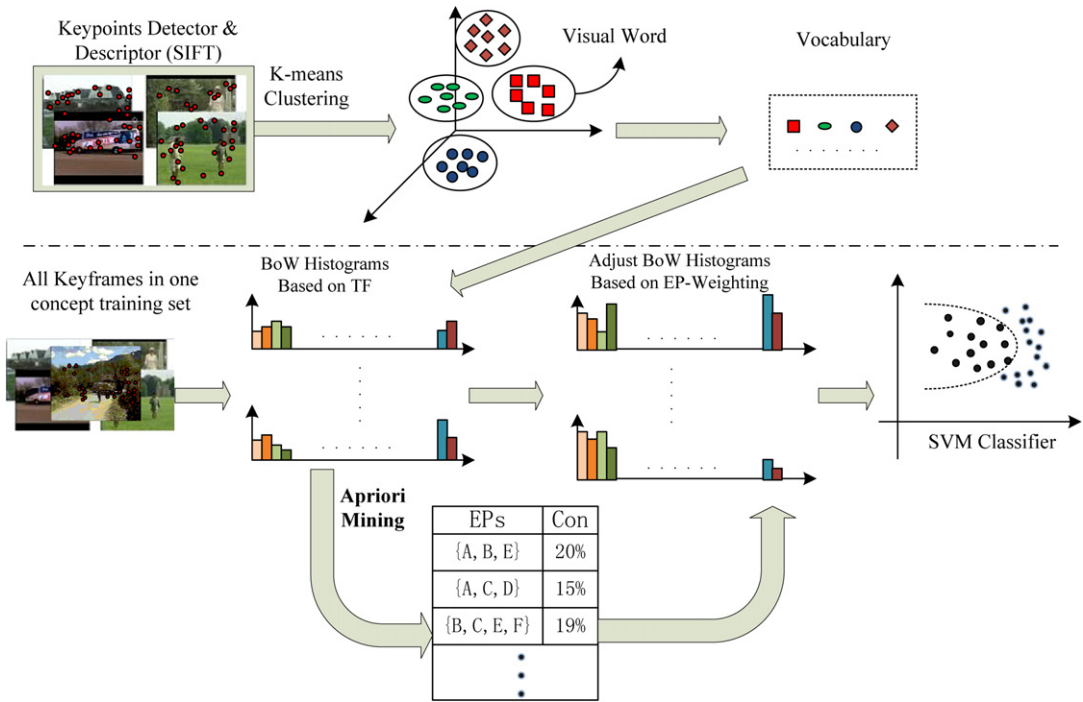


Fig. 2. Architecture of the proposed scheme.

this paper, we investigate how to utilize EPs to adjust the importance of each visual word in BoW.

3. The proposed visual word weighting scheme

In this paper, we introduce EPs mining into video annotation, and present a novel visual word weighting scheme. The architecture of the proposed scheme is illustrated in Fig. 2. The main steps are as follows. Firstly, the local features are extracted from salient image patches by SIFT method [2]. Secondly, a visual vocabulary is constructed through a k-means clustering algorithm to cluster the local keypoints. Each keypoint cluster is treated as a visual word in the visual vocabulary. Thirdly, by mapping the keypoints in an image to the visual vocabulary based on TF, an image can be described as a feature vector (referred as BoW histogram). Fourthly, through Apriori algorithm, the EPs are mined from the feature vector set for all keyframes in one single concept training set. Finally, the confidence values of EPs are used to adjust the BoW histograms obtained in the third step. And the adjusted BoW histograms are used as the training data of SVM classifier to acquire the classifier for each semantic concept. The details of each step will be described in the following subsections.

3.1. Mining emerging patterns

In this section, we describe the EPs mining process. Given a set of n training images $T = \{I_1, I_2, \dots, I_n\}$, let $F = \{F_1, F_2, \dots, F_n\}$ be a set of features of all images and $C = \{C_1, C_2, \dots, C_k\}$ be a set of k possible semantic concepts contained in the images. As defined above, a training image I_i can be represented as $I_i = (I_i \cdot F_i, I_i \cdot C_i)$, where

$I_i \cdot F_i$ is the feature of I_i and $I_i \cdot C_i$ is the semantic concept of I_i . Each concept C_i in T can be divided into two subsets: the positive set $T_{C_i}^+$ and the negative set $T_{C_i}^-$. To easily explain the EPs mining, the following definitions are given:

Definition 1. Given a pattern P , the support $Sup_p(C_i)$ and confidence $Con_p(C_i)$ for a concept C_i are defined as

$$Sup_p(C_i) = \frac{num(T_{C_i}^+ + T_{C_i}^-, P)}{num(T_{C_i}^+ + T_{C_i}^-)}$$

$$Con_p(C_i) = \frac{num(T_{C_i}^+, P)}{num(T_{C_i}^+, P) + num(T_{C_i}^-, P)}$$

where $num(T, P)$ represents the number of images that contain pattern P in set T , $num(T)$ represents the number of all images in set T .

Definition 2. Given min_sup and min_con , a pattern P is the emerging pattern (EP) for a concept C_i , if it satisfies

$$Sup_p(C_i) \geq min_sup, \quad Con_p(C_i) \geq min_con$$

Based on the definitions above, the improved Apriori algorithm [4] is used to mine EPs. The process of mining EPs is summarized in Algorithm 1.

3.2. EP-weighting scheme

According to the confidence of each EP and how many concepts it occurs in, we design the EP-weighting scheme to adjust the visual word weights given by TF scheme.

Algorithm 1. Mining Emerging Patterns

Input: Training Dataset: $T_{C_i}^+$ and $T_{C_i}^-$ for concept C_i , the thresholds: min_sup and min_con

Output: Emerging Patterns for concept C_i : $EP(C_i)$

- (1) Call Apriori ($T_{C_i}^+ + T_{C_i}^-$) to collect the set of frequent patterns and their supports, and then select the pattern set satisfying $\text{Sup}_p(C_i) \geq \text{min_sup}$ as candidate pattern set $CPs(C_i)$.
- (2) Calculate the confidence of every pattern in $CPs(C_i)$ and select the pattern set satisfying $\text{Con}_p(C_i) \geq \text{min_con}$ as Emerging Patterns Set $EPs(C_i)$.
- (3) Output $EPs(C_i)$.

Without loss of generality, let $EPs(C)$ be the set of all EPs in the set of concepts C , P_j be an emerging pattern in $EPs(C)$. The process of calculating the important factor is firstly depicted as follows.

Algorithm 2. Calculating the important factor α for every P_j

Input: all emerging patterns P_j in $EPs(C)$

Output: the important factors for pattern P_j

```

for each emerging pattern  $P_j$  in  $EPs(C)$ 
   $L = 0$ 
  for each  $C_i$  in  $C$ 
    if  $P_j$  is emerging for  $C_i$ 
       $L = L + 1$ ;
    end if
  end for
   $\alpha = 1/L$ 
  output  $\alpha$  for pattern  $P_j$ 
end for

```

From Algorithm 2, we can know that, if $\alpha = 1$, P_j is only emerging in some one concept C_i . Then P_j is very decisive for concept C_i . Therefore, during training SVM classifier, the weight of P_j should be larger than those EPs occurring several times in other concepts. According to the important factor α and the confidence of emerging pattern, the process of calculating the weighting factor ω of each emerging pattern in concept C_i is summarized in Algorithm 3.

Algorithm 3. Calculating the weighting factor ω of each P_j in concept C_i

Input: all emerging patterns P_j in $EPs(C_i)$, their confidence and adjustment factor α

Output: the weighting factors of pattern P_j for concept C_i

```

for each emerging pattern  $P_j$  in  $EPs(C_i)$ 
   $\omega_{P_j} = e^{\alpha \cdot \text{Con}_{P_j}(C_i)}$ 
  output  $\omega_{P_j}$  of  $P_j$  for concept  $C_i$ 
end for

```

The weighting factor is used directly to adjust the weights of BoW. Let W be a vocabulary of n visual words: $W = (w_1, w_2, \dots, w_n)$. With the vocabulary, a training or testing image I for concept C_i can be described as a feature vector $F_I = (v_{w_1}, v_{w_2}, \dots, v_{w_n})$, where v_{w_1} represents the feature value of word w_1 in the image, which

Algorithm 4. Adjusting the feature values of visual words

Input: the feature vector of a training or testing image for concept C_i : $F_I = (v_{w_1}, v_{w_2}, \dots, v_{w_n})$, the weighting factor of pattern P_j for C_i

Output: the adjusted feature vector

```

for each visual word  $w_i$ 
  for each emerging pattern  $P_j$  in  $EPs(C_i)$ 
    if  $w_i \in P_j$ 
       $v_{w_i} = \omega_{P_j} \times v_{w_i}$ ;
    end if
  end for
  output  $v_{w_i}$ 
end for

```

is calculated by TF scheme. Based on the weighting factors calculated by Algorithm 3, the feature values of visual words are adjusted as Algorithm 4.

4. Experimental results

To evaluate the proposed scheme for video annotation, we conduct the experiments on the benchmark video corpus of the TRECVID 2006 dataset, which consists of 137 broadcast news videos [8]. These training videos are segmented into 61,901 video shots and 39 concepts are labelled as positive ("1") or negative ("0") on each shot. In the experiments, 20 concepts are selected, and the key-points are detected by DoG detector and described by the PCA-SIFT descriptor. We adopt the k-means clustering algorithm to generate the visual word vocabulary. In [3], Y.G. Jiang et al. had proven the impact of vocabulary size was insignificant when one more sophisticated weighting scheme was employed. The performances of BoW were similar on TRECVID dataset for vocabulary sizes ranging from 500 to 10,000. Therefore, we generate a vocabulary containing 2000 visual words for efficiency. With the vocabulary, 20 binary SVM classifiers [14] are trained for these selected semantic concepts, where each classifier is for determining the presence of a specific concept. 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. We average the APs over all the 20 concepts to create the Mean Average Precision (MAP), which is the overall evaluation result.

In [5], five existing weighting schemes were evaluated. They had proven the un-normalized scheme were always better than their normalized counterparts for TRECVID dataset. Therefore, in the experiments, we focus on the comparison among the proposed EP-weighting scheme, the binary scheme, the TF scheme, and the TF-IDF scheme. For all the key frames, the BoW features are respectively calculated based on the four weighting schemes of visual-word feature vectors. For each setting, the linear and RBF kernel of SVM are used, and the performance of the better one is reported. The two parameters in Algorithm 1 are empirically chose as $\text{min_sup} = 0.15$ and $\text{min_con} = 0.25$. Fig. 3 illustrates the AP of the four schemes. We can see that the EP-weighting scheme outperforms the other schemes for

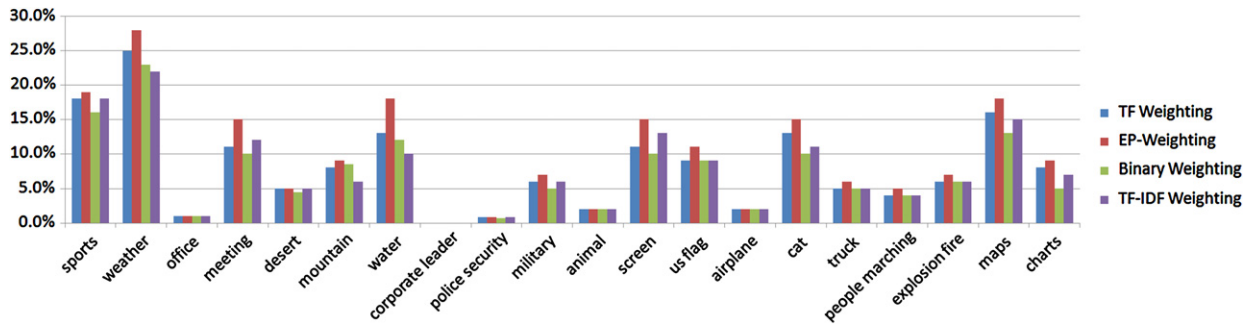


Fig. 3. Performances of the four weighting schemes over 20 concepts.

Table 1

MAP of the four weighting schemes over 20 concepts.

Weighting schemes	Binary weighting	TF weighting	TF-IDF weighting	EP weighting
MAP	7.3%	8.2%	7.7%	9.6%

over 14 of all 20 concepts. Some of the improvements are significant, such as the 36%, 36%, and 38% improvements on “meeting”, “screen” and “water”, respectively. The proposed scheme remains with no change on a few concepts. The main reason is that these concepts contain a much smaller amount of EPs, which fails to adjust the weights of visual words.

Table 1 summarizes the MAP performances of the four weighting schemes. The EP-weighting scheme achieves around 17% improvements compared to the TF scheme. These results demonstrate that it is effective that the EPs are used to adjust the importance of visual words in BoW for video annotation.

5. Conclusions

In this paper, we have presented a novel scheme for weighting BoW feature based on EPs. By mining the EPs of visual words for each concept, the effects of EPs' visual word are improved in concept classifiers. The new scheme can efficiently capture the co-occurrence relationships of visual words. Experiments on the benchmark TRECVID data set demonstrated that the proposed scheme was superior to the existing weighting schemes. For future work, we will focus on two directions. One direction investigates how to utilize the language model in visual word description. One possibility is to treat an image as a paper contained many visual words and explore the use of text retrieval technologies for video classification. Another direction is to improve annotation performance by leveraging on position information of each visual word.

Acknowledgements

The authors acknowledge the support received from the National Natural Science Foundation of China (Project

60972096) and National 863 Plans Projects (Grant No. 2009AA01Z410).

References

- [1] H.T. Shen, B.C. Ooi, X. Zhou, Z. Huang, Towards effective indexing for large video sequence data, in: SIGMOD 2005, Baltimore, USA.
- [2] D. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* 60 (2) (2004) 91–110.
- [3] Y.G. Jiang, C.W. Ngo, J. Yang, Towards optimal bag-of-features for object categorization and semantic video retrieval, in: ACM International Conference on Image and Video Retrieval (CIVR'07), Amsterdam, Netherlands, Jul. 9–11, 2007.
- [4] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: International Conference on Computer Vision, 2003.
- [5] J. Yang, Y.G. Jiang, A.G. Hauptmann, C.W. Ngo, Evaluating Bag-of-Visual-Words representations in scene classification, in: ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'07), Augsburg, Germany, Sep. 2007.
- [6] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, *International Journal of Computer Vision* 60 (1) (2004) 63–86.
- [7] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (10) (2005) 1615–1630.
- [8] <http://www-nlpir.nist.gov/projects/trecvid/>.
- [9] S.F. Chang, J. He, Y.G. Jiang, et al., Columbia University VIREO-CityU IIRIT TRECVID2008 high-level feature extraction and interactive video search, in: NIST TRECVID Workshop (TRECVID'08), Gaithersburg, MD, USA, Nov. 17–18, 2008.
- [10] Tao Mei, Zheng-Jun Zha, et al., MSRA at TRECVID 2008 high-level feature extraction and automatic search, in: NIST TRECVID Workshop (TRECVID'08), Gaithersburg, MD, USA, Nov. 17–18, 2008.
- [11] G. Dong, J. Li, Efficient mining of emerging patterns: Discovering trends and differences, in: Proc. 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99), San Diego, CA, USA, Aug. 1999, pp. 43–52.
- [12] J. Li, G. Dong, K. Ramamohanarao, Making use of the most expressive jumping emerging patterns for classification, *Knowledge Information System* 3 (2) (2001) 131–145.
- [13] H. Fan, K. Ramamohanarao, A weighting scheme based on emerging patterns for weighted support vector machines, in: IEEE International Conference on Granular Computing, Beijing, China, July 2005, pp. 435–440.
- [14] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery* 2 (2) (1998) 121–167.
- [15] Y.G. Jiang, C.W. Ngo, Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval, *Computer Vision and Image Understanding* 113 (3) (2009) 405–414.