DISCRIMINATIVE BAG-OF-VISUAL PHRASE LEARNING FOR LANDMARK RECOGNITION

Tao Chen, Kim-Hui Yap and Dajiang Zhang
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
{Chen0523, ekhyap}@ntu.edu.sg, Dzhang3@e.ntu.edu.sg

ABSTRACT

Bag-of-visual phrase (BoP) has been proposed and developed for landmark recognition recently. However, existing BoP methods for landmark recognition have two major shortcomings: (i) they try to construct a universal phrase vocabulary for all object categories, which lacks specific descriptive capabilities for a particular category, and (ii) they often adopt simple criterion such as the frequency information to mine the visual phrases, which may cause the selected phrases to be less discriminative or representative for recognition. In view of this, this paper proposes a new discriminative BoP approach for landmark recognition. First, the candidate visual phrases defined as adjacent pairwise words are selected for each category. A phrase-level similarity measure at the latent space is proposed to evaluate the semantic similarity between pairwise phrases. This is then integrated with the phrase frequency information to shortlist the discriminative phrases for each category through a proposed phrase ranking algorithm. Finally, the BoP and bag-of-words (BoW) histograms are combined through a pyramid matching method for recognition. Experimental results on two different datasets demonstrate that the proposed method is effective in landmark recognition.

Index Terms-BoW, BoP, discriminative visual phrases, landmark recognition

1. INTRODUCTION

In recent years, bags-of-words (BoW) methods have been widely used in various landmark/place recognition systems [1]-[4] and has demonstrated good performance [5]. However, one shortcoming of BoW is that it assumes the local features in an image are independent from each other. As a result, the generated visual words are also independent and their contextual relationships are ignored, which is important for human beings to understand an image. Considering this, some efforts have been put into mining the contextual relationship between the local features and in using a set of visual phrases to represent an image [6]-[9]. They concentrate on developing various methods to mine useful visual phrases for recognition. A simple method to discover visual phrase that comprises pairwise words is proposed in [6]. The occurrence frequencies of various words are utilized to determine visual phrases. A more general method that tries to mine visual phrases containing 2, 3, and 4 words is proposed in [7]-[8]. Both of them first use *K*-nearest neighbors to construct a word-set database, and then employ data mining techniques to discover important visual phrases. A contextual BoW method is proposed in [9]. Semantically similar and spatial neighboring words are mined simultaneously to form various visual phrases. They are then used to generate a histogram through quantization operation.

The visual phrase methods have obtained good performance in some applications. However, a drawback still exists in their endeavors, that is, they focus on mining the shared phrases that appear frequently in most categories. This causes the discovered phrases to lack descriptive abilities for a specific semantic category. Although some preliminary efforts to learn the descriptive visual phrases for each category have been made in [6], it is far from enough since it only considers the frequency information and ignores (i) the meaningless background phrases and (ii) the different semantic similarity between pairwise phrases.

In order to overcome these problems, this paper proposes a new discriminative bags-of-phrase (BoP) approach to complement the conventional BoW and BoP methods for landmark recognition based on category-based discriminative visual phrase mining. The proposed approach mainly consists of two components:

(i) A visual phrase-level semantic similarity measure at the latent space based on the PLSA and Bayesian estimation. Kullback-Leibler (KL) divergence is utilized to measure the semantic distance between two phrases.

(ii) A visual phrase ranking algorithm that takes advantage of phrase-level semantic similarity and frequency information in order to discover the descriptive phrases for each category. The selected phrases of each category are then combined to form the final vocabulary.

The rest of this paper is organized as follows. The discriminative BoP learning technique is discussed in Section 2, which includes the proposal of the semantic similarity measure between pairwise phrases and phrase ranking algorithm. Experiment results and discussions are given in Section 3. Section 4 concludes the paper with a summary of our findings.

2. DISCRIMINATIVE VISUAL PHRASE MINING

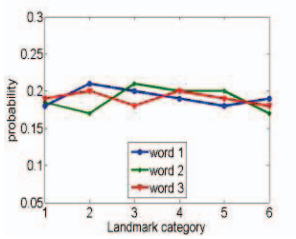
In this section, we will present the method to mine the discriminative visual phrases for each landmark category. SIFT keypoints are first detected in each image and the descriptors are clustered to generate a codebook. We first define a set of notations: (i) visual word codebook: $\Omega = \{w_1, ..., w_M\}$, where M is the codebook size; (ii) Second-

order word-sets: $V=\{v_1,v_2,...,v_N\}$ where $v_n=\{w_i,w_j\}$ is generated using the method in [6], which uses a spatial histogram to detect the word-sets in a predefined neighborhood. The word-sets that have larger occurrence frequencies are selected as visual phrase candidates. A phrase ranking algorithm that integrates the phrase-level semantic similarity and frequency information is proposed to discover the discriminative phrases for each category.

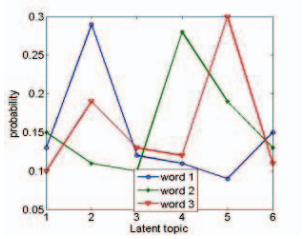
2.1 Semantic similarity between pairwise phrases

In order to measure the phrase-level semantic similarity, we should first measure the word-level similarity, and then extend it to phrase-level similarity. Previous work utilizes the category distribution induced by visual words to measure the semantic distance between various words [9]. The category c distribution conditioned on the words w, denoted as $P(c|w)$, can be interpreted as how much the word votes for each of the categories whenever it occurs. However, one condition for this method to perform well is that the images of a category should not contain similar objects that appear in other categories, which in fact is hard to guarantee. To demonstrate this, Fig. 1(a) shows several images from different landmark categories. We can see that the occurrence probabilities of the word related to the concept of "window" (denoted as "word 1") in these categories are nearly the same and has minor distribution differences among various categories. Similarly, the background words related to the concept "greenery" and "ground" (denoted as "word 2" and "word 3" in Fig. 1(b) and (c)) also span across many categories and have similar category distribution with the "word 1". Therefore, these three words may be falsely grouped together.

(a)



(b)



(c)

Fig.1 (a) Sample images from 6 landmark categories, (b) Similar category distributions induced by three words, (c) Distinct latent topic distributions induced by three words

Considering this problem, we propose to model a latent topic distribution induced by visual words, which is denoted as $P(z|w)$, where z is latent topics. Similarly, $P(z|w)$ can be interpreted as how much the word w votes for each of the latent topics. The intuition behind this approach is that visual words that belong to the same semantic object are more likely to have similar distribution over the latent topics than over the categories. The reason is that a latent topic is more concrete in representing a specific semantic concept while a category may contain multiple semantic objects. In this work, we use PLSA [11] to infer the latent topics. We randomly sample a number of images from each category to form the dataset, and represent each image by a BoW histogram. Suppose $P(d)$ denotes the probability of observing an image d in the dataset, $P(w|z)$ denotes the conditional probability of a word w conditioned on the latent topic z, and $P(z|d)$ denotes an image specific probability distribution over the latent space. PLSA model is then generated using the following three steps: (i) Choose an image d with probability $P(d)$, (ii) pick a latent topic z with probability $P(z|d)$, and (iii) generate a word w with probability $P(w|z)$. As a result, we will obtain an observation pair (w, d). Repeat this process for several times, we will obtain a co-occurrence matrix $n(w,d)$. The parameters $P(z)$, $P(w|z)$ and $P(d|z)$ are determined by maximizing the log-likelihood function:

$$L=\log P(D,W)=\sum_{d\in D}\sum_{w\in W} n(w,d)\log P(w,d)$$
$$s.t.\sum_{z\in Z}P(z)=1,\ \sum_{w\in W}P(w|z)=1,\ \sum_{d\in D}P(d|z)=1 \quad (1)$$

The model is fitted using the Expectation Maximization (EM) algorithm as described in [11]. Bayesian estimation is then adopted to infer the latent topic distribution induced by the visual words, denoted as $P(z|w)$,

$$P(z|w)=\frac{P(z,w)}{P(w)}=\frac{P(w|z)P(z)}{\sum_{z\in Z}P(w|z)P(z)} \quad (2)$$

Fig. 1(c) illustrates the latent topic distributions induced by three words (6 latent topics are modeled in PLSA here), which shows that the three distributions have significant differences, and thus can be easily distinguished. Using (2), the semantic distance between two words w_i and w_j can be measured by a weighted average of the Kullback–Leibler (KL) divergence of each latent topic distribution to their mean distribution, which can overcome the asymmetric shortcoming of single KL divergence:

$$d(w_i,w_j)=\frac{P(w_i)}{P(w_i)+p(w_j)}\cdot KL\left(P(z|w_i)\,\|\,P(z|w_i\cup w_j)\right)$$
$$+\frac{P(w_j)}{P(w_i)+P(w_j)}\cdot KL\left(P(z|w_j)\,\|\,P(z|w_i\cup w_j)\right) \quad (3)$$

where $P(z|w_i\cup w_j)$ is the mean average of $P(z|w_i)$, $P(z|w_j)$. The KL divergence is defined as,

$$KL\left(P(z|w_i)\,\|\,P(z|w_i\cup w_j)\right)=\sum_{z\in Z}P(z|w_i)\log\frac{P(z|w_i)}{P(z|w_j\cup w_j)} \quad (4)$$

Using (3), the distance between phrases p_a, p_b is defined as:

$$d(p_a, p_b) = \min_R d_R(p_a, p_b)$$
$$d_R(p_a, p_b) = \sum_{\substack{w_i \in p_a, w_j \in p_b \\ R: w_i \to w_j}} d(w_i, w_j) \quad (5)$$

where $R: w_i \to w_j$ is the match order from the words in p_a to that in p_b [10]. There are two possible match orders for second-order phrases in this work. The match order that produces the smaller distance between the two phrases is adopted. Finally, we use a normalization function to convert the pairwise phrase distance value to a semantic similarity value within [0, 1] as follows:

$$s(p_a, p_b) = \ln\left(1 + \frac{1}{d(p_a, p_b)}\right) \quad (6)$$

2.2 Visual phrase ranking algorithm

In this section, we propose a visual phrase ranking algorithm that leverages the idea of word ranking in [6] to select the descriptive visual phrases (DVP) for each category. A matrix **L** is constructed to record (i) the phrase discrimination for its category, which is indicated by the phrase frequency ratio of the positive to negative categories in the diagonal element **L**(*i, i*), and (ii) the phrase representative capability for its category, which is indicated by its semantic similarity with other phrases in the off-diagonal elements **L**(*i, j*). The diagonal and off-diagonal elements of the matrix are defined as:

$$L(i,i) = P(p_i)\ln\left(1 + \frac{P(p_i)}{\overline{P}(p_i)}\right),\ L(i,j) = s(p_i, p_j) \quad (7)$$

where P, $\overline{P}$ correspond to the occurrence probabilities of phrase p in the positive and negative categories. After computing the matrix, it is normalized by column to 1. We set the initial rank value of each candidate phrase to be equal to 1 and then start the rank-updating iteration by multiplying the matrix **L** with the rank vector *r* as:

$$r^{(n+1)} = \mathbf{L}\, r^{(n)} \quad (8)$$

where (*n*) denotes the *n*-th iteration. During the iteration, the candidates having greater inherent importance and stronger semantic relationship with weighted candidates will be ranked higher. Iterations are carried out to update the weight of each phrase until the weight converges. A certain number of phrases with large ranking values in *r* are selected as the DVP. The DVPs of each category are then combined to form a universal BoP vocabulary to encode detected local feature sets in each image into a BoP histogram.

Finally, in order to utilize the respective advantages of BoW and BoP histograms (denoted as X_{BoW} and X_{BoP}) for recognition, the pyramid matching method in [4] is adopted to fuse the two histograms into one vector. χ^2 function is used as the matching kernel. The new pyramid matching kernel for two images' histograms *X* and *Y* is defined as:

$$K(X,Y) = \alpha\frac{(X_{BoW} - Y_{BoW})^2}{X_{BoW} + Y_{BoW}} + \beta\frac{(X_{BoP} - Y_{BoP})^2}{X_{BoP} + Y_{BoP}} \quad (9)$$

where X and Y are the combination of BoW and BoP histograms, and can be represented as $X = X_{BoW} \cup X_{BoP}$, $Y = Y_{BoW} \cup Y_{BoP}$, α, β are the order weights and can be set as 0.25, 0.5, respectively by cross validation, similar to [4].

3. EXPERIMENTAL EVALUATION

In this experiment, a landmark database consisting of 3622 training images and 534 testing images using 50 categories of landmarks from the campus in Nanyang Technological University (NTU) is created. Landmark is defined as a building or place-of-interest that is unique or distinctive. Fig. 2 shows some sample images of the 50 landmarks. For each landmark category, there are on average 70 images for training, and 10 images for testing. The images are captured using camera phones under different capturing conditions, including scale, illumination, viewpoint and color changes, etc. All images are resized to 320×240 or 240×320 pixels. Support vector machine is used as the classifier.

Fig. 2 Sample images of 50 landmark categories

Table 1 Performance comparison of BoW, proposed BoP, and their combination

Vocabulary size / Recognition rate	300	600	900	1200	1500
BoW (%)	78.1	81.3	82.7	82.3	81.0
Proposed BoP (%)	78.3	82.1	84.5	87.3	88.8
BoW + BoP (%)	79.0	83.1	86.0	88.8	90.4

The performance comparison of conventional BoW, proposed BoP and their combination for landmark recognition are given in Table 1. The vocabulary sizes of BoW and BoP range from hundreds to thousands. The vocabulary size of BoW and BoP fusion is increased when combined. Table 1 shows that the performance of the proposed BoP consistently outperforms the conventional BoW method for each group of vocabulary size. This can be explained by the fact that visual phrase can better describe the words spatial contextual relationship than individual

words. Further, it is noted that the BoW and BoP combination for recognition outperforms the BoP or BoW alone. This can be attributed to the fact that BoW and BoP can complement each other by describing not only the image's first-order word distribution but also second-order spatial distribution. Specifically, a highest recognition rate of 90.4 % is obtained when combining BoW and BoP for the vocabulary size of 1500.

Next, comparison experiments between the proposed and existing BoP methods [6][9] are conducted. The results are given in Fig. 3. From the figure, it can be seen that the proposed BoP approach consistently outperforms the other two visual phrase methods, especially when the vocabulary size is increased to larger range. This shows the effectiveness of the proposed BoP approach.

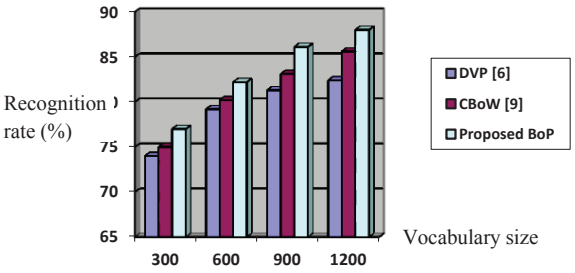


Fig.3 Performance comparison of the proposed BoP method against the contextual BoW [9] and DVP [6] methods.

In order to demonstrate the effectiveness of the proposed method on other datasets, the Oxford building dataset that consists of 5062 images is downloaded from [12]. It is manually annotated for 11 different landmarks. Each landmark contains 5 queries. The comparison results are given in the Table 2. It is worth mentioning that the recognition rate of the BoP method on Oxford database can be boosted by increasing the size of the BoP vocabulary. Here we only provide the performance comparison for the BoP vocabulary size of 1200. From the table, it can be seen that the proposed method achieves a good recognition accuracy of 81.8%, which is 9.5%, and 3.6%, higher than the method in [6] and [9] respectively. This shows that the proposed method can obtain good performance in a different benchmark dataset. Further, the reason that the performance (81.8%) on the Oxford dataset is lower than that (87.3%) on the NTU dataset is due to different database collection process. The NTU dataset is collected for training purpose and includes different capturing conditions, while Oxford dataset is collected for retrieval purpose.

Table 2 Experimental results on the Oxford building dataset

Method	Recognition accuracy (%)
CBoW method [9]	78.2
DVP [6]	72.3
Proposed method	81.8

4. CONCLUSION

This paper presents a new discriminative BoP approach for landmark recognition based on latent-space visual phrase learning. The key contributions of this paper include: (i) proposal of a latent space discriminative phrase selection approach, which utilizes the PLSA and Bayesian estimation to select the descriptive phrases for each category, and (ii) development of an effective phrase ranking algorithm, which integrates the phrase-level frequency ratio and semantic similarity for DVP mining. Experimental results on two datasets show that the proposed method can achieve good recognition performance in landmark recognition.

5. REFERENCES

[1] Y. P. Li, D. J. Crandall, and D. P. Huttenlocher, "Landmark classification in large-scale image collections," *Proceedings of IEEE International Conference on Computer Vision*, Kyoto, Japan, 2009.

[2] T. Chen, K.- H. Yap, L.-P. Chau, "A discriminative learning approach for mobile landmark recognition," *IEEE International Conference on Image Processing*, pp. 217-220, Brussels, Belgium, Sep. 2011.

[3] T. Chen, K. –H. Yap, and L. –P. Chau**,** "Integrated content and context analysis for mobile landmark recognition," *IEEE Transactions on Circuits and Systems for Video Technology***,** vol. 21**,** pp. 1476 - 1486, 2011.

[4] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169-2178, 2006.

[5] K.–H. Yap, T. Chen, Z. Li, K. Wu, "A comparative study of mobile-based landmark recognition techniques," *IEEE Intelligent Systems*, vol. 25, no. 1, pp. 48-57, Jan./Feb. 2010.

[6] S. L. Zhang, Q. Tian, G. Hua, et al., "Descriptive visual words and visual phrases for image applications," *ACM International Conference on Multimedia*, 2009.

[7] J. Yuan, Y. Wu, M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[8] Y.-T. Zheng, M. Zhao, S.-Y. Neo, et al., "Visual synset: Towards a higher-level visual representation," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2008.

[9] T. Li, T. Mei, I. –S. Kweon, X.-S. Hua, "Contextual bags-of-words for visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2010.

[10] Q. Tian, S. Zhang, W. Zhou, R., Ji, B., Ni, and N. Sebe, "Building descriptive and discriminative visual codebook for large-scale image applications," *International Journal of Multimedia Tools and Applications,* 2010.

[11] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 41, no. 2, pp. 177-196, 2001.

[12] The Oxford Visual Geometry Research Group: www.robots.ox.ac.uk/~vgg/data/oxbuildings/index.html.