A Fast Method of Visual Words Assignment of Bag-of-Features for Object Recognition

Meng-Jiun Chiou[†1] Toshihiko Yamasaki[†2] Kiyoharu Aizawa[†3]

1. Introduction

This paper deals with object recognition and retrieval problems, especially for large-scale pictures categorization. For these large-scale object recognition problems, high-speed processing is an important factor with respect to practicality. Extracting the features of query pictures is one of the most common processes for state-of-the-art strategies. However, it has been pointed out that feature extraction is one the most time-consuming parts in the object recognition problem [1]. To deal with this problem, this paper introduces a fast look-up table based method of finding bag-of-features-based indexes of query pictures, without conducting feature extraction. The proposed look-up table consists of local patches (stored in a pixel form) and their corresponding visual words (assigned from the visual words codebook) of images. By this, the visual word is assigned to the query by retrieving the most similar patch in the database. In our experiment, we compared the proposed method with other common feature extraction methods. In addition, our experiment processes inherits MATLAB code of ScSPM [2], while not using sparse coding but just amended to evaluate the proposed method.

2. Background

In this chapter, we introduce the overall process of object recognition used in our experiment. The process consist of three parts: 1) SIFT [3] (feature extraction) 2) Bag-of-features [4] (representation model) and 3) Classification.

2.1 SIFT

In order to describe a picture in a simple way, feature extraction is used in most common solutions. SIFT (Scale-Invariant Feature Transform), is used for extracting scale-invariant keypoints and computing its descriptor to form distinctive image features. In our experiment, SIFT is mainly computed by two steps: 1) Keypoints Detection: For standard SIFT, images across various scale are convolved with Gaussian filter, and keypoints are detected by Difference of Gaussian (DoG). While in our experiment, instead of using DoG, SIFT from 16×16 pixel patches are densely sampled on a grid with the stepsize of 8 pixels. 2) Feature Description: The direction and magnitude of gradient are computed for every pixel in a neighboring region of the detected keypoint. Finally, the neighboring region are divided into 16 blocks of 4×4 pixel, each with 8 bin orientation vector, which form an SIFT of 128-dimensional vector.

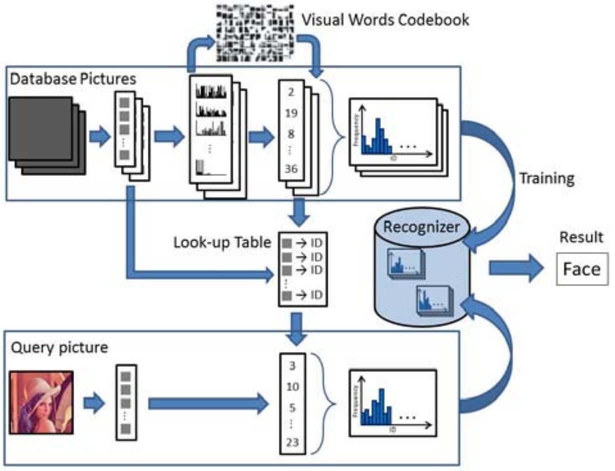


Fig. 1: The flow chart of proposed Method

2.2 Bag-of-features

Bag-of-features is a popular representation model, usually used to represent a single picture. First, feature extraction (e.g. SIFT) is implemented. Next, extracted features are used to train a visual words codebook through K-means clustering, then the nearest visual word is assigned to the extracted features from the visual words codebook by vector quantization. All assigned visual words of a single picture form a high dimensional (which depends on number of visual words) histogram vector.

2.3 Classification

Such vectors described above can be trained through classifiers (e.g. Support Vector Machine) into a database, which can be used to predict the class of query pictures. Also, query pictures are in the same representation as database pictures (i.e. histogram vectors here). In this step, the accuracy of prediction will be evaluated across different sizes of training and table-forming database.

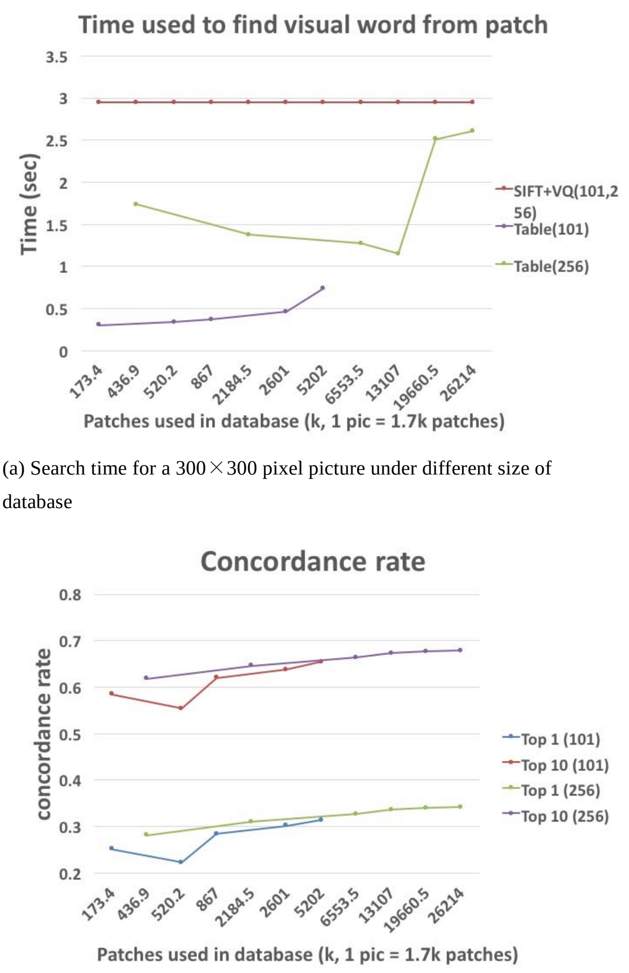
3. Bag-of-Features with Look-Up Table

To deal with the problem of a large number of query tables, we applied look-up table to Bag-of-Features. Look-up table is composed of: 1) 16×16 pixel patches of database pictures densely sampled on a grid (stored in pixel form) at feature extraction step, and 2) their corresponding visual words assigned from visual words codebook. As shown in Fig. 1, by look-up table, instead of extracting SIFT, query pictures are split

[†1] Exchange Student at The Faculty of Engineering, The University of Tokyo, and Degree Student at National Chiao Tung University, Taiwan

[†2] Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo

[†3] Department of Information and Communication Engineering, The Faculty of Engineering, The University of Tokyo



(a) Search time for a 300×300 pixel picture under different size of database

(b) Concordance rate (top-1 and top-10) under different size of database

Fig. 2: Experiment result on Caltech101/Caltech256

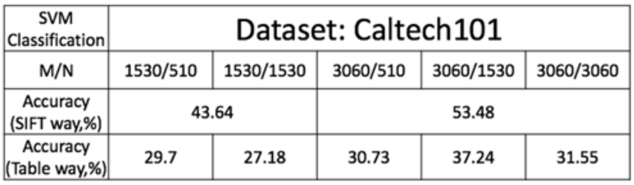
into 16×16 pixel patches. Then, these patches are used to search the nearest patches in database so that its corresponding visual words got. To ensure the precision, how to find the nearest patch is important.

4. Experiment

To evaluate the proposed method, we extended the look-up table from only one nearest visual word stored, into 10 of that stored so that we can calculate concordance of them. Top-1 means that the nearest visual word found by FLANN is the same as the nearest one found by SIFT way, while top-10 means the former is included in the top 10 nearest visual word found by SIFT way. Also, we implemented common process feature extraction and visual words assigning on the query pictures. Therefore, we could calculate the concordance between the visual words found by look-up table and the visual words assigned by codebook.

We adopted FLANN [5] as our indexing and searching method to find the nearest patch, and used dataset Caltech101 [6] and Caltech256 [7]. In addition, the pictures were all preprocessed into gray scale and normalized (mean is zero). This experiment was implemented at a server with 24 processors (Intel® Xeon® E5-2643 v2 @ 3.5 GHz) and 512 GB memories.

Table 1: Accuracy of proposed method compared to SIFT way. M, N is the number of pictures used to train database and form table, respectively.



SVM Classification	Dataset: Caltech101				
M/N	1530/510	1530/1530	3060/510	3060/1530	3060/3060
Accuracy (SIFT way,%)	43.64		53.48		
Accuracy (Table way,%)	29.7	27.18	30.73	37.24	31.55

As shown in Fig. 2 (a), the proposed method is faster compared to common SIFT method under various size of database for both dataset. Time of assigning visual word to pictures is reduced by up to 85% in condition of small database. While in Fig. 2 (b) showed that the top-1 concordance of patches is about 25-35% (the nearest visual word found by FLANN and the answer visual word found by SIFT) and the top-10 concordance is about 55-70%. Also, a similar pattern showed on both databases Caltech101/Caltech256.

To realize the performance of the proposed method, classification based on histogram vectors, which is summation of visual words formed by top-1 visual word, was implemented by Support Vector Classification (RBF kernel) of LIBSVM [8]. As shown in Table 2, speedy as the proposed way is, it achieved lower accuracy compared to SIFT way.

5. Conclusion

This study presented a high-speed way to deal with large-scale image classification. The proposed look-up table created by local patches and their corresponding visual words is used to finding bag-of-features-based indexes of query pictures. With the proposed method, time of assigning visual word can be reduced by up to 85%, while lower accuracy achieved compared to common solution. The proposed method needs to be evaluated under more conditions to verify its practicability.

Reference

[1] Wu, Y., Lu, S., Mei, T., Zhang, J., & Li, S. "Local visual words coding for low bit rate mobile visual search." in *Proceedings of the 20th ACM international conference on Multimedia* (pp. 989-992), 2012.
[2] Yang, Jianchao, et al. "Linear spatial pyramid matching using sparse coding for image classification," in CVPR 2009.
[3] D. G. Lowe, "Distinctive image features from scale-invariant key-points," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[4] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 1470–1477.
[5] Marius Muja and David G. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration", in VISAPP 2009.
[6] L. Fei-Fei, R. Fergus and P. Perona. "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories", in CVPR 2004.
[7] Griffin, G. Holub, and P. AD. Perona. "Caltech-256 object category dataset." Technical Report 7694, California Institute of Technology, 2007.
[8] Chih-Chung Chang and Chih-Jen Lin. "libsvm: a library for support vector machines". ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, 2011.