Spatial Rerank-based Bag of Words Model

Thuyen V. Phan
Advanced Program in Computer Science
Faculty of Information Technology
University of Science, VNU - HCM
Email: pvthuyen@apcs.vn

Minh-Bao Truong
Advanced Program in Computer Science
Faculty of Information Technology
University of Science, VNU - HCM
Email: tmbao@apcs.vn

***Abstract*—The popularity of many social network sites such as Facebook, Twitter, and YouTube creates a huge demand of managing and querying visual data from billions of images. Searching by text cannot describe the information as effective as by images but it is more complicated to search by images due to the changes of camera angles or lighting conditions. To address this problem, the authors conduct a comparison of the retrieval quality between a standard Bag-of-Words model and one with spatial reranking using RANSAC. Our experimental results on Oxford Building 5K Dataset show that spatial reranking improves the mean Average Precision of Bag of Words model from 0.708 up to 0.742. The authors view this work as a promising step to further improve the performance of different visual search systems using the spatial relations between the images.**

I. Introduction

Alongside the intensive growth of social networks such as Facebook or Twitter, the information that users want to share is not only text but also other complex types, especially images. In January 2009, Kaplan et al. state that there are over 3 billion photos on Flickr [1]. And in 2012, it is recorded that 250 million photos are uploaded to Facebook everyday [2]. This fact creates a huge demand of managing and querying data from huge amount of information in different formats (text, image, video, sound ...).

The tendency to search not only by text but also some special features of other complex types such as images is currently one of the most popular concerns. Thus, huge search engines such as Google and Bing have already developed and integrated their visual search system using images as queries. Nevertheless, the problem of improving performance of visual search systems remains an interest of many research lab and corporation.

To query visual data using a single image, one of many approaches is template matching method, i.e. a technique for finding small parts of an image which match a template image [3], [4], [5]. Another popular technique is to evaluate the similarity of two images by comparing some regions which seem to be the interested points of the images, namely features matching [6], [7], [8]. The algorithm that the authors choose to discuss in this paper is Bag-of-Words (BoW) [9] which is used by many different image retrieval systems [9], [10], [11]. A reason why Bag-of-Words is widely used is that its allow parts of a query image to appear flexible in the result images. Hence, BoW model is a really potential approach and is focused by many research groups.

However, due to the flexibility among parts of the query images, BoW might not fully exploit the spatial relations among the components of images. This fact motivates our investigation to prove that cooperating the spatial information can increase the precision of BoW model. Our experiments performed on Oxford Building 5K Dataset shows that using an extra spatial rerank step has a huge impact on the BoW model in term of mean Average Precision (an increase from 0.708 to 0.742). This work's main contribution is the proof that spatial information is really useful when retrieving visual information and promote the research of enforcing spatial consistency to image retrieval systems.

The rest of this paper is organized as follows. In section II, we review the background and related words in image retrieval and image classification. The core steps of the BoW model and how we conduct experiments are presented in section III. Section IV is for experiment results and evaluations. The conclusion and future works are presented in section V.

II. Background & Related works

There are many approaches to create Image Information Retrieval System. Some methods aim at high precision, i.e achieve high quality of top retrieved results, others focus on high recall, i.e retrieve all positive results. Among them, the first effective and scalable method is Bag-of-Words, Sivic and Zisserman [9], which is inspired by the correspondence algorithm using in text retrieval. Before going into details of BoW model in subsection II-B, we will first introduce some different methods for image retrieval problem in subsection II-A.

A. Different approaches for image retrieval problem

One of many popular method is histogram comparisons which compares 2 different images based on their color histograms. Some early works of this approach using a cross-bin matching cost for histogram comparison can be found in [12], [13], [14]. In [14], Peleg et al. represent images as sets of pebbles after normalization. The similarity score is then computed as the matching cost of two sets of pebbles based on their distances.

Another well-known technique is template matching, i.e. seeking a given pattern in a image by comparing to candidate regions of the same size in the target image. By consider both the pattern and candidate regions as a length-N vector, we can

compare these two vectors by using different kind of distance metrics, one such metric is the Minkowski distance [15].

B. Bag-of-Words

Since Bag-of-Words is originally a text retrieval algorithm, we will first introduce some backgrounds about BoW in text retrieval problem in subsection II-B1 before discussing about using BoW in image retrieval in subsection II-B2.

1) Bag-of-Words in text retrieval: In text retrieval, a text is represented as a histogram of words, also known as BoW [16]. This scheme is called term frequency weighting as the value of each histogram bin is equal to the number of times the word appears in the document. Moreover, some words are less informative than others since those words appear in almost every document. Therefore, we need a weighting scheme that address this problem. Such weighting scheme is called inverse document frequency (idf) and is formulated as $log(N_D/N_i)$, where N_D is the number of documents in the collection and N_i is the number of documents which contains word i. The overall BoW representation is thus weighted by multiplying the term frequency (tf) with the inverse document frequency (idf) giving rise to the tf-idf weighting [16]. Extremely frequent words, "stop words", can be removed entirely in order to reduce storage requirements and query time.

2) Bag-of-Words in image retrieval: When applying BoW to image retrieval, a major obstacle is the fact that text documents are naturally broken into words by spaces, dots, hyphens, or commas. In contrast, there is no such separator in images. Therefore, the concept of "visual word" is introduced where each visual word is represented as a cluster obtained using k-means on the local descriptor vectors [9].

The bigger the vocabulary size is, the more different the visual words are. Hence, the vocabulary helps us distinguish the images more effectively. Nonetheless, with bigger vocabulary size, slightly different descriptors can be assigned to different visual words thus not contributing to the similarity of the respective images and causing a drop in performance examined in [17], [18], [11]. Philbin et al. [11] suggests "soft assign" method where each descriptor is assigned to multiple nearest visual words instead of using "hard assignment", i.e only assign a local descriptor to only one nearest visual word. Despite its effectiveness, this method also significantly costs more storage and time.

III. Method

A. Overview of an Image Retrieval System

As described in Figure 1, a typical Image Retrieval System consists of 2 main steps: Feature Extraction and Similarity Matching. We will explain how we do the Feature Extraction step in this section. In section III-B, the authors focus on describing how we perform Similarity Matching with our implementation of BoW model.

To extract features from images, the authors choose to use Hessian affine detector [19]. As we tested on Oxford Building 5K Dataset, there are typically 3,300 features for each image and a total about 16 millions of features for the whole dataset.

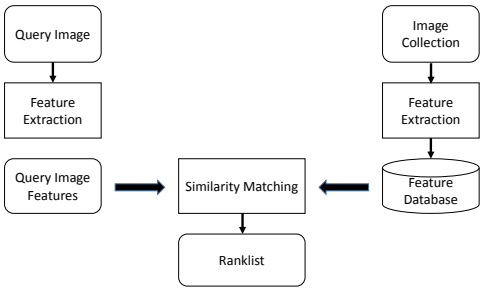

Fig. 1. How an image retrieval system works

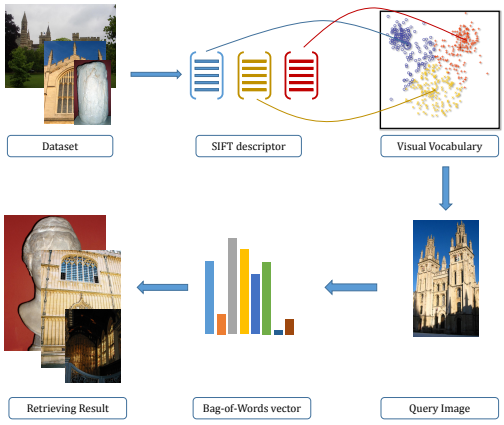

Fig. 2. Diagram of Bag-of-words model

Then, we compute the SIFT descriptors [20] of all the features and use these descriptors for the next steps.

B. Similarity Matching

As showed in Figure 2 typical BoW model used in image retrieval systems would consist of the following steps: Dictionary Building, Quantization and Retrieving the result. The 2 former steps are described in section III-B1 and section **??**. The last step is our main focus in this paper and is discussed in section III-B3 and section III-B4.

1) Dictionary Building: Generating the dictionary of visual words for such a huge amount descriptors is a big challenge. In order to overcome this obstacle, the authors use the approximate k-means (AKM) instead of the traditional exact k-means (KM). AKM is proposed by Philbin et al. [10], which reduce the majority amount of time taken by exact nearest neighbors computation by using an approximate method instead. Also, in [10], Philbin et al. shows that using 1M dictionary size would have the best performance on the Oxford Building 5K Dataset.

2) Quantization: Subsequently, each 128-dimension SIFT descriptor is reduced to a 3-dimension vectors of their 3 nearest visual words in the dictionary. Each of these 3 nearest cluster is assigned with weights calculated with the formula proposed by Sivic et al. [11], $weight = \exp(-\frac{d^2}{2\delta^2})$, where d is the distance from the cluster center to the descriptor point and δ^2 is chosen to be 6250. Then, by adding all these weights to

their corresponding bins, we will have the BoW representation of an image.

3) tf-idf Weighting Scheme: As mentioned in section II, tf-idf is a popular weighting scheme that is used by almost any BoW model. In this section, the authors will show how this scheme is applied to our system.

For a term t_i in a particular document d_j, its term frequency $tf_{i,j}$ is defined as follow:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

Where $n_{i,j}$ is the number of occurrences of the considered term t_i in the document d_j. The denominator is the sum of the number of occurrences of all the terms in document d_j.

The inverse document frequency idf_i of a term t_i is computed by the following formula:

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2}$$

Where, $|D|$ is the total number of documents in the corpus, $|\{j : t_i \in d_j\}|$ is the number of documents where the term t_i appears, i.e. $n_{i,j} \neq 0$

The tf-idf weight of a term t_i in a document d_j is then calculated as the product of tf and idf:

$$tfidf_{i,j} = tfi, j \times idf_i \tag{3}$$

The tf-idf weight is then used to compute the similarity score between an image d_i and a query q:

$$s_{d_i,q} = t\vec{fidf}_i \cdot t\vec{fidf}_q = \sum_{j=1}^{|T|} tfidf_{i,j} \times tfidf_{q,j} \tag{4}$$

Finally, by sorting the list of images corresponding to their similarity score with a query, we achieve the raw ranked list of this query which is then used for the Spatial Rerank step.

4) Spatial Rerank: So far, we have always consider an image as a document of visual words, which means that we totally ignore the spatial structure of the features. Thus, we now incorporate the spatial constraints to the top ranked result and rerank them. The spatial verification process evaluate a geometry transformation based on features coordinates of a image and the query. The target images are then reranked using the sum of the spatial verified visual words' idf. A common approach is to use RANSAC [21], i.e. generating different transformation hypotheses and choose the hypothese which has the largest number of "inliers". In our system, we perform spatial rerank on the top 800 retrieved result of the dataset, which is showed to obtain the best accuracy by Philbin et al. [10].

IV. Experiment & Result

To prove our hypotheses, the authors test both the BoW systems with and without spatial rerank on the Oxford Building 5K Dataset [22]. This dataset was constructed by Philbin et al. in 2007 [10]. It consists of 5,062 images of resolution 1024×768 belongs to 11 different Oxford buildings. Images for each building are collected from Flickr by searching using text queries. In figure 3, some samples from the dataset are shown. To evaluate our system, the authors use 55 available queries along with their ground-truth, 5 for each landmark, as shown in figure 4.

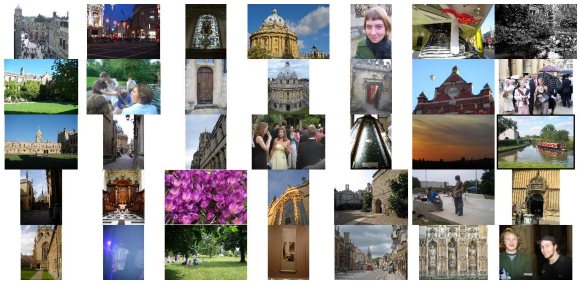

Fig. 3. Some random images from Oxford Building 5K Dataset

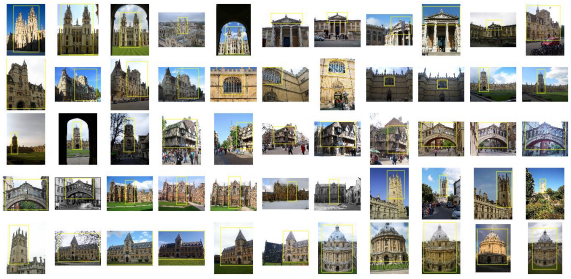

Fig. 4. 55 queries of Oxford Building 5K Dataset

References

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business Horizons*, vol. 53, no. 1, pp. 59 – 68, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0007681309001232

[2] R. E. Wilson, S. D. Gosling, and L. T. Graham, "A review of facebook research in the social sciences," *Perspectives on Psychological Science*, vol. 7, no. 3, pp. 203–220, May 2012.

[3] R. Brunelli, *Template Matching Techniques in Computer Vision: Theory and Practice*. Wiley, 2009.

[4] A. Rosenfeld and G. J. Vanderburg, "Coarse-fine template matching," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 7, no. 2, pp. 104–107, Feb 1977.

[5] M. Gharavi-Alkhansari, "A fast globally optimal algorithm for template matching using low-resolution pruning," *Image Processing, IEEE Transactions on*, vol. 10, no. 4, pp. 526–533, Apr 2001.

[6] S. Belongie, C. Carson, H. Greenspan, and J. Malik, "Color- and texture-based image segmentation using em and its application to content-based image retrieval," in *Computer Vision, 1998. Sixth International Conference on*, Jan 1998, pp. 675–682.

[7] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[9] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, 2003.

[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. CVPR*, 2007.

[11] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. CVPR*, 2008.

[12] H. C. Shen and A. K. Wong, "Generalized texture representation and metric," *Computer Vision, Graphics, and Image Processing*, vol. 23, no. 2, pp. 187 – 206, 1983. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0734189X83901123

[13] M. Werman, S. Peleg, and A. Rosenfeld, "A distance metric for multidimensional histograms," *Computer Vision, Graphics, and Image Processing*, vol. 32, no. 3, pp. 328 – 336, 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0734189X85900556

[14] S. Peleg, M. Werman, and H. Rom, "A unified approach to the change of resolution: space and gray-level," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, no. 7, pp. 739–742, Jul 1989.

[15] W. Ouyang, F. Tombari, S. Mattoccia, L. Di Stefano, and W.-K. Cham, "Performance evaluation of full search equivalent pattern matching algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 127–143, Jan 2012.

[16] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[17] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, 2006.

[18] G. Schindler, M. Brown, , and R. Szeliski, "City-scale location recognition," in *Proc. CVPR*, 2007.

[19] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[20] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[21] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981. [Online]. Available: http://doi.acm.org/10.1145/358669.358692

[22] [Online]. Available: http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/