VISUAL-WORD-BASED DUPLICATE IMAGE SEARCH WITH PSEUDO-RELEVANCE FEEDBACK

Jen-Hao Hsiao[1,2], Chu-Song Chen[2,3], and Ming-Syan Chen[1]

1 Dept. of Electrical Engineering, National Taiwan University, Taipei, Taiwan.
2 Institute of Information Science, Academia Sinica, Taipei, Taiwan.
3 Gradual Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan.

ABSTRACT

We aim to improve the bag-of-visual-words (BOW) model for near-duplicate image retrieval, by introducing a more fine-grained pseudo-relevance feedback process. The BOW method is based on vector quantization of affine invariant descriptors of image patches. Despite its popularity and simplicity, the retrieval performance of BOW is often unsatisfactory due to the large and diverse variations of near-duplicate images. We thus propose an information-theoretic feedback framework that employs available cues in the search result to find more relevant duplicate images which are hard to retrieve by using conventional BOW approaches. Our algorithm is experimentally evaluated under a severely attacked image database, and shown to significantly improve the retrieval accuracy over a non-feedback baseline.

Index Terms — Image retrieval, Near-duplicate image retrieval, Pseudo relevance feedback.

1. INTRODUCTION

Advancement in the Internet technology has led to the growth of multimedia information, such as digital images publishing on Web sites. As content owners need to trace the uses of their images and to prevent the copyright infringement, finding copies in a large image database has become an important issue.

Near-duplicate image retrieval is one of such technique that can find similar images with certain variations induced by some common image manipulations, such as image enhancements and geometric transformation. One leading method [7] for image object retrieval and near-duplicate image detection from large image database was inspired by the bag-of-words approach in text information retrieval. In the bag-of-words representation, a text document is encoded as a histogram of the number of occurrences of each word. Similarly, one can characterize an image by a histogram of visual words count called bag-of-visual-words (BOW), and the image retrieval task can then work as text information retrieval. The main difference to text retrieval is that there is no pre-given visual vocabulary for the image retrieval problem, and thus they have to be learned from a training set in advance.

However, due to the large and diverse variations of near-duplicate images, results of the BOW methods are often unsatisfactory. The reason of the limited performance can be summarized as two-folds. First, by replacing invariant descriptors with prototypes, such approaches discard some discriminating content present in the unique features, and consequently tend to suffer more acutely from false negatives than most text retrieval systems. Second, image retrieval differs from common tasks of classification such as face detection and character recognition. In a retrieval task, the number of potential image classes is extremely large and usually only a small number (or even only one) of query examples are given. The insufficient information of the user's query induces the term-mismatch problem, which causes the retrieval system to provide non-relevant images to the user, and spends lots of time on reformulating queries until relevant information are retrieved. Consequently, poor queries often bring up unsatisfactory results. It is why we need approaches for leveraging the available cues in the search set of BOW retrieval without putting extra burden on the user or system developers.

Relevance feedback is an effective way to gather information about the query class distribution and has long been used to enhance retrieval performance for image retrieval systems. The general idea is to use the images identified by the user as relevance to modify the current query so that in the next search iteration a more effective query can be used. Through the interaction between the retrieval system and user, the system is expected to provide more accurate retrieval results.

Motivated by the concept of relevance feedback, in this paper, we focus on improving the BOW retrieval model. The relevance-based language model is used as our basic retrieval framework, allowing more evidences to be incorporated into the feedback procedure. Since it is impossible to obtain user judgments in automatic retrieval tasks, the pseudo-relevance-feedback (PRF) technique [2] that assumes the cues in top-ranked documents are relevant to query is adopted in our feedback process.

The remainder of this paper is organized as follows. Section 2 reviews the related researches. Section 3 describes the proposed framework, including the language model and

the pseudo-relevance feedback approach. Section 4 presents the experimental results of the retrieval accuracy of our approach. Finally, conclusions are given in Sections 5.

2. RELATED WORKS

The idea of visual words was initially proposed in [7]. In this work, Sivic and Zisserman performed retrieval of shots from a movie using a text retrieval approach. However, the retrieval/classification performance of BOW is not totally satisfied. Many methods have thus proposed to further improve the effectiveness, including visual phrase-based approach [10], visual cue cluster [3], and language model-based approach [9]. However, due to the large and diverse variations in near-duplicate images, retrieving near-duplicate images with high performance still remains difficult.

3. PROPOSED FRAMEWORK

Below we detail the proposed framework, including the BOW, language model, and pseudo-relevance feedback strategy.

3.1. Bag-of-Visual-Words Representation

The basic idea of BOW is to treat images as a collection of the representative prototypes sampled from training image corpus, and then use the resulted distribution in the descriptor space as a characterization of the image.

In this paper, the SIFT algorithm [5] is employed to extract the feature descriptors due to its impressive performance in image recognition. The resulted 128-dimensional vector of edge orientation histogram (EOH) constructed by SIFT is employed as the distinctive local descriptor.

To build the visual vocabulary, one of the important challenges is the scalability of construction method. Since the size of visual vocabulary is an important parameter that can affect the system performance, a proper number of visual words is benefit to balancing the retrieval accuracy and response time. Howver, conventional BOW approaches adopted the k-means algorithm that is easy to implement but hard to scale to a large vocabulary. To improve the scalability in visual vocabulary construction, we employed a clustering approach based on approximate-nearest-neighbor (ANN) [1] to balance the computation time and quantization accuracy.

As we know that the computation effort in typical k-means is mostly spent on calculating the nearest neighbors between the points and cluster centers. We have thus tried to lessen the computation by replacing the exact nearest-neighbor (NN) problem with an approximate k-d tree (Ak-d tree) [1]. ANN provides a speedup over the exact nearest neighbor by several orders of magnitude yet results in only slight accuracy degradation. Therefore, replacing this exact computation in k-means by an ANN method lessens the computation effort, and enhances the scalability of visual vocabulary. In this case, the Ak-d tree is built over the cluster centers at the beginning of each iteration of clustering process to increase the training speed.

Finally, query and database images are all mapped into a histogram of visual words. Similarly, Ak-d tree can also be employed to accelerate the mapping of image descriptors to BOW in the on-line retrieval step.

3.2. Relevance-based Language Model for Retrieval

We use a language modeling-based approach as our retrieval framework. The basic idea of this retrieval model is to measure the distribution similarity (or relevance value) between the query model and document model by employing KL divergence [4]. Hence, the retrieval problem turns out to be an estimation problem that estimates the unigram models for a query and a set of documents.

More formally, suppose that we are given a collection **T** of target (database) images. Each image $I \in \mathbf{T}$ is represented by a discrete set of visual words, $F(I) = \{w_1, w_2, \ldots, w_n\}$, generated as described in Section 3.1. Assume that a database image I is obtained as a sample from a unigram language model (i.e., a multinomial word distribution) $P(w|\ \theta_I)$ with parameters θ_I. The simplest way to estimate the document language model is to treat the image as a sample from the underlying multinomial word distribution and use the maximum likelihood estimator [4]:

$$P(w \mid \hat{\theta}_I) = \frac{tf(w, I)}{|I|} \tag{1}$$

where $tf(w, I)$ is the count of the visual word w in image I, and $|I|$ is the total number of words in I.

However, (1) could generate a zero probability if a visual word never occurs in the document image I, causing problems in scoring the likelihood of a document with the query. To avoid the incorrectness, Dirichlet smoothing technique is employed. Dirichlet smoothing uses a document-dependent coefficient (parameterized with μ) to control the interpolation,

$$P(w \mid \hat{\theta}_I) = \frac{tf(w, I) + \mu P(w \mid \theta_b)}{|I| + \mu}. \tag{2}$$

Here $P(w|\theta_b)$ is the probability of visual word w given by the collection language model θ_b, which is usually estimated by using the whole collection of image documents **T**, e.g.

$$P(w \mid \theta_b) = \frac{\sum_{I \in T} tf(w, I)}{\sum_{I \in T} |I|}. \tag{3}$$

The generative model for a query is simply set as a unigram language model, which is defined as:

$$P(w \mid \hat{\theta}_Q) = P(w \mid Q) = \frac{tf(w, Q)}{|Q|}, \tag{4}$$

where Q denotes the query image, $tf(w, Q)$ is the count of the visual word w in query Q, and $|Q|$ is the total number of visual words in Q.

Given the estimated query and a document language model $\hat{\theta}_Q$ and $\hat{\theta}_I$, the relevance value of I with respect to Q can then be measured by the KL-divergence:

$$KLD(\hat{\theta}_Q \| \hat{\theta}_I) = \sum_{w \in V} P(w|\hat{\theta}_Q) \log \frac{P(w|\hat{\theta}_Q)}{P(w|\hat{\theta}_I)}, \quad (5)$$

where **V** is the set of all the visual words.

3.3. Pseudo Relevance Feedback

The KL-divergence-based retrieval method described in the previous section, however, is not very discriminative because the information contained in a single query image is limited. In this section, we explore the Pseudo-Relevance Feedback (PRF) [6] technique that can further improve the estimation of θ_Q by further exploiting the information contained in top-ranked retrieval images.

PRF aims to enhance the retrieval performance by generating query expansion keywords from the target collection (e.g., the evaluation database). Based on the assumption that a query image is generated by sampling from two different language models: a query image model and a feedback query model, the language model of a new query $\hat{\theta}_Q^t$ can be defined as a linear interpolation of these models,

$$P(w|\hat{\theta}_Q^t) = \lambda P(w|\hat{\theta}_Q) + (1-\lambda) P(w|\hat{\theta}_F), \quad (6)$$

where λ is a parameter controlling the influence of the feedback model, and $P(w|\hat{\theta}_F)$ denotes the estimated feedback query model based on the feedback images, which are the top-ranked images from the initial search result.

A natural way to estimate a feedback query model $\hat{\theta}_F$ is to assume that the feedback documents are generated by a probabilistic model $P(w|\hat{\theta}_F)$, which is a summation over language models of the top-ranked search images (denoted by R):

$$P(w|\hat{\theta}_F) = \frac{1}{|R|} \sum_{I \in R} P(w|\hat{\theta}_I) KLD(\hat{\theta}_Q \| \hat{\theta}_I). \quad (7)$$

However, PRF is an unsupervised learning process that can not guarantee whether the pseudo-relevant images are truly relevant or not. In this case, if the quality of the initial search is extremely low, then the new query model could be worse than the original one because the noisy information is inevitably integrated into the feedback process. These noises can cause a serious problem from a practical point of view that the system responses a messed-up ranked output to users after a two-stage search.

To prevent the false positives in top-ranked images from being joined into the feedback process, we have thus performed a post-verification stage to verify the usefulness in these feedback candidate images.

The post-verification proceeds by matching individual SIFT keypoint of the query image to each of the images in the top-ranked list in the descriptors space. When the distance ratio (i.e., the closest neighbor to that of the second-closest neighbor when matching two images) is smaller than a threshold τ, the image in the top-ranked list gets a vote. Finally, only the verified image set, R^*, having the number of votes above a threshold η will join the feedback procedure. Here we sets τ = 0.7 and η = 5 as suggest in [5] to filter out the irrelevant feedback candidate images

4. EXPERIMENTS AND OBSERVATIONS

We start this section by describing the experimental setup and implementation details of the proposed framework. We then conduct several experiments to show the effectiveness of our approach.

4.1. Experimental Setup

We use Corel image library that has been widely used for image retrieval and copy detection as our base data set. We generate the testing images (duplicate images) by randomly selected 2,000 images from the base set, and then transformed each of them into 18 modified versions by using StirMark 4.0 [8], a standard benchmark originally designed to evaluate the robustness of digital watermark. The image attacks used in the experiment include affine transformation, JPEG quantization, noise, convolution, scaling, rotation, median filtering, cropping, and self-similarity (i.e., color space transformation). Totally, we have generated 2,000×18=36,000 testing images to server as target (database) collection.

The retrieval performance is evaluated with the probability of the successful top-k retrieval [9], defined as:

$$S\_Prob(top-k) = \frac{Q_r}{Q_A}, \quad (8)$$

where Q_r is the number of duplicate images found in the top-k list, and Q_A is the total number of duplicate images. We experimentally select 3000 visual word as the size of the visual vocabulary, and use μ=10 and λ=0.5 as the parameters of the language model for all the following experiments.

4.2. Results

We first examine the improvement of retrieval performance by combining the language model and pseudo-relevance feedback. The results are summarized as the curves shown in Figure 1. The curves marked by 'BOW-KLD' and 'PRF' are both obtained by using the KL-divergence retrieval model, but 'PRF' has further performed a single-round pseudo-relevance feedback. Note that in this case, only verified feedback candidate images are used in the feedback process. As can be seen, the PRF-based approach achieves better retrieval performance than pure BOW for all the different top-k values. This reveals that the proposed approach can effectively improve the duplicate image retrieval performance through pseudo-relevance feedback. We also compare language-model-based KL-divergence with some other distance metrics. The curves marked by 'BOW-ECL' and 'BOW-COS' are obtained by using the

histogram of visual words count as the image signature, but employing Euclidean and cosine distance respectively as distance metrics. As can be seen, KL-divergence achieves a better retrieval performance among these three metrics.

Figure 2 has further shown the effect of post-verification. Figure 2(a) shows the retrieval performance without verifying the feedback candidate images. It is obvious that PRF without verifications have unstable performances because the noises cause degradation on the estimation of the feedback model. The probabilities of successful retrieval are even worse when the joined feedback images and the number of feedback rounds (FBs) increase. On the contrary, Figure 2(b) shows the retrieval performance of PRF with post-verification. As can be seen, the probabilities of successful retrieval are stably increased with the number of feedback rounds, and gradually converge after several rounds of feedback. The post-verification for PRF can thus effectively filter out irrelevant information.

5. CONCLUSIONS

In this paper, we have presented an image retrieval method by integrating the language-modeling framework with PRF. The top-ranked images retrieved are post-verified to join the feedback process so that the performance degradation caused by irrelevant noise can be eliminated. The proposed PRF framework has shown its improvements on the retrieval performance than the pure BOW retrieval framework.

ACKNOWLEDGEMENT

This research was supported in part by Taiwan E-learning and Digital Archives Programs (TELDAP) sponsored by the National Science Council of Taiwan under NSC Grants: NSC 96-3113-H-001-010, NSC 96-3113-H-001-011 and NSC 96-3113-H-001-012.

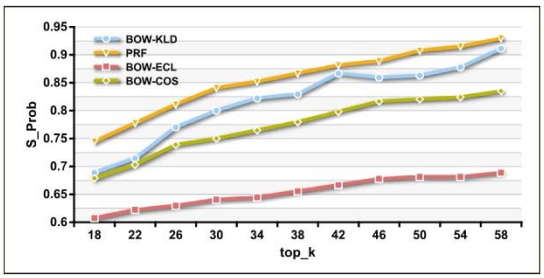
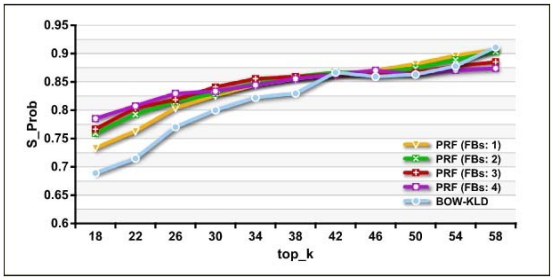
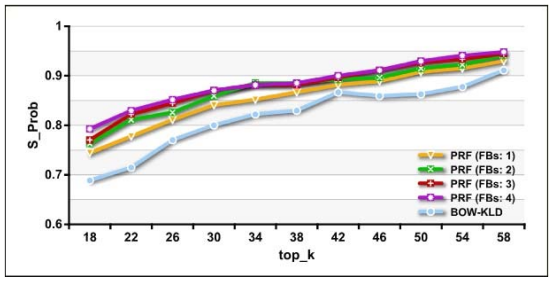


Figure 1. Performance comparison of different retrieval approaches.



(a)



(b)

Figure 2. Effect of post-verification. (a) PRF with all feedback images. (b) PRF with verified feedback images only.

6. REFERENCES

[1] S. Arya, D.M. Mount, N. S. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching. *Journal of the ACM*, vol. 45, pp. 891-923, 1998.

[2] J. G. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee. Translingual information retrieval: A comparative evaluation. *International Joint Conference on Artificial Intelligence*, pp. 708-715, 1997.

[3] W. H. Hsu and S.-F. Chang. Visual cue cluster construction via information bottleneck principle and kernel density estimation. *Proceedings of the International Conference on Content-Based Image and Video Retrieval*, pp. 82-91, 2005.

[4] J. Lafferty and C. Zhai, Document language models, query models, and risk minimization for information retrieval, Proceedings of the international *ACM SIGIR conference on Research and development in information retrieval*, pp. 111-119, 2001.

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.

[6] J. G. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee. Translingual information retrieval: A comparative evaluation. *International Joint Conference on Artificial Intelligence*, pp. 708-715, 1997.

[7] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. *Proceedings of IEEE International Conference on Computer Vision*, pp. 1470- 1477, 2003.

[8] Stirmark benchmark 4.0, available at http://www.petitcolas.net/fabien/watermarking/stirmark/.

[9] X. Wu, W.-L. Zhao, and C.-W. Ngo, Near-duplicate keyframe retrieval with visual keywords and semantic context, *Proceedings of ACM international conference on Image and video retrieval*. pp. 162-169, 2007.

[10] Q.-F. Zheng, W.-Q. Wang, W. Gao, Effective and efficient object-based image retrieval using visual phrases, *Proceedings of ACM international conference on Multimedia*, pp. 77-80, 2006.