Large-scale Image Retrieval based on the Vocabulary Tree

Bo Cheng, Li Zhuo, Pei Zhang and Jing Zhang
Signal & Information Processing Laboratory, Beijing University of Technology, Beijing, China

Keywords: Vocabulary Tree, Large-scale Image Retrieval, Optimized SIFT, Local Fisher Discriminant Analysis.

Abstract: In this paper, vocabulary tree based large-scale image retrieval scheme is proposed that can achieve higher accuracy and speed. The novelty of this paper can be summarized as follows. First, because traditional Scale Invariant Feature Transform (SIFT) descriptors are excessively concentrated in some areas of images, the extraction process of SIFT features is optimized to reduce the number. Then, combined with optimized-SIFT, color histogram in Hue, Saturation, Value (HSV) color space is extracted to be another image feature. Moreover, Local Fisher Discriminant Analysis (LFDA) is applied to reduce the dimension of SIFT and color features, which will help to shorten feature-clustering time. Finally, dimension-reduced features are used to generate vocabulary trees which will be used for large-scale image retrieval. The experimental results on several image datasets show that, the proposed method can achieve satisfying retrieval precision.

1 INTRODUCTION

Image retrieval has been an active research topic in recent years due to its potentially large impact on both image utilization and organization. Researchers aim to seek ways that have greater promptness and accuracy.

Content-Based Image Retrieval (CBIR) is currently considered as the mainstream method because of desirable processing speed and objectivity. It detects and extracts visual features of image (e.g. global feature and local feature) automatically by means of image processing and computer vision algorithm. In most of cases, a retrieval system takes visual features of a query image given by a user, and then the features are compared with the features stored in a database. As a result, the user will receive images that have similar features with the query image.

CBIR mainly includes two key parts: feature extraction and similarity comparison. The features usually can be divided into two kinds: global features and local features. The most commonly used local features contain Scale Invariant Feature Transform (SIFT, Lowe D. G., 2004), Principle Component Analysis-SIFT (PCA-SIFT, Ke Y. et al., 2004), Speeded Up Robust Features (SURF, Bay H. et al., 2008), and Gradient Location-Orientation Histogram (GLOH, Mikolajczyk K. et al., 2005) as well. Relying on grey information of images, SIFT features which are adopted to operate accurate and speedy image retrieval from large-scale database can be invariant to changes of image scaling, rotation, illumination, and others. PCA-SIFT performs well in terms of image rotation, blur and illumination changes, while not at scaling and affine transformations. Moreover, projection matrix of PCA-SIFT needs a series of typical images, which is only appropriate for the specific type. SURF runs three times faster than SIFT concerning computational complexity. It also works more robust than SIFT when blurred images are processed. Nonetheless, SURF does not operate as well as SIFT in dealing with images affected by scaling, rotation and illumination changes. As the extension of SIFT, GLOH can improve robustness and discrimination performance of the descriptors.

Establishing an effective index mechanism is another critical aspect to fulfill fast retrieval in the large-scale image database. Currently, there are three kinds of methods: K-D Tree (Böhm C. et al., 2001), LSH (Gionis A. et al., 1999), and vocabulary tree (Nist´er D. et al., 2006). K-D tree uses the nearest neighbour search to build the index of the images. Its search accuracy is higher in low dimensional space while the performance of K-D tree drops rapidly when dimensions are increasing. LSH can be used to reduce dimensions with multiple

Cheng B., Zhuo L., Zhang P. and Zhang J..
Large-scale Image Retrieval based on the Vocabulary Tree.
DOI: 10.5220/0004661802990304
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 299-304
ISBN: 978-989-758-004-8

hash functions which encode in a low-dimensional space to represent the higher one. However, as code length and number of images grow, this method can neither improve query accuracy significantly, nor offer fair image retrieval performance because a large number of codes require vast storage space. Compared with the methods mentioned above, vocabulary tree can effectively shorten matching time and improve retrieval performance. The method based on feature distribution selects the appropriate cluster centre by integrating different clustering to generate a tree structure of feature classification.

In this paper, a large-scale image retrieval method based on vocabulary tree is proposed. The novelty of this paper consists of three main aspects. First, traditional SIFT descriptors are excessively concentrated in some areas of images, the extraction process is optimized to reduce the number of SIFT features. Then, combined with optimized-SIFT, color histogram in Hue, Saturation, Value (HSV) color space is extracted as a global feature to represent image content. Moreover, Local Fisher Discriminant Analysis (LFDA, Rahulamathavan Y. et al., 2013) is applied to reduce the dimension of SIFT and color features, which will shorten feature-clustering time. Finally, dimension-reduced features are used to generate vocabulary tree which will be used for large-scale image retrieval. By comparing multiple sets of experimental data, it can be concluded that the proposed method can achieve satisfying retrieval performance.

The rest of this paper is organised as follows: section 2 introduces the proposed large-scale image retrieval framework based on vocabulary tree. Section 3 presents the experimental results. Final conclusions are drawn in section 4.

2 LARGE-SCALE IMAGE RETRIEVAL FRAMEWORK BASED ON VOCABULARY TREE

The proposed large-scale image retrieval framework based on vocabulary tree is shown in Figure 1. First, features are respectively extracted from image database. Then, these features whose dimension will be reduced are used to construct two vocabulary trees by means of hierarchical *K*-means clustering scheme. By counting inverted index (Zobel J. and Moat A., 1998) based on vocabulary tree, the features and their indexes are stored in the image database.

When operating a query, the features of query image are extracted and indexed as former. The similarity of images is measured and ranked by comparing the index of the query image with those stored in the database. The most similar top-k images will be regarded as research results and returned to the user.

As Figure 1 shows, the proposed scheme includes four segments: feature extraction, dimensionality reduction, the construction of the inverted index of vocabulary tree and similarity measurement. They will be further introduced below.

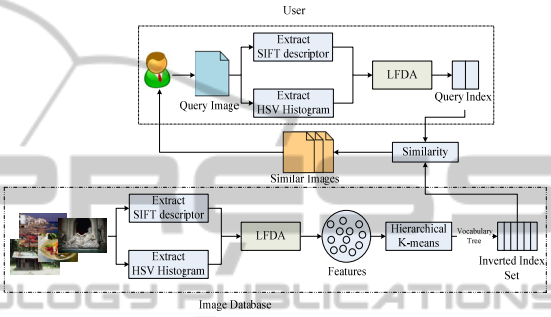

Figure 1: Large-scale image retrieval framework based on vocabulary tree.

2.1 Feature Extraction

It is clear that feature extraction is one of the most critical parts of the image retrieval framework. In this paper, SIFT and color features are extracted as image features. SIFT extraction procedure is properly optimized to reduce excessive number of SIFT features.

2.1.1 SIFT Features

Traditional SIFT extraction is shown in Figure 2(a), where the crossings present the determined coordinates of SIFT features.

It can be found that SIFT descriptors are over-concentrated in the areas with similar characteristics. Therefore, we propose a method to optimize SIFT feature extraction procedure to reduce the number of SIFT features. After optimization, the image content can be still characterized accurately, but with fewer SIFT descriptors.

Suppose that $Sift_{des}[i].x$, $Sift_{des}[i].y$ respectively represents the horizontal and vertical coordinates of i-th SIFT, T_{opt} is an optimization threshold, and R_{opt} is optimization range. For any two different SIFT descriptors $Sift_{des}[i]$ and $Sift_{des}[j]$, when the distance of horizontal and vertical coordinates of two points are less than the optimal threshold T_{opt}, these two

(a)Traditional SIFT （b）Optimized SIFT

Figure 2: SIFT descriptors.

points can be merged into one because they exist in the optimization range R_{opt}. Otherwise, it is needless to optimize. The optimization process can be represented in Equation (1):

$$Sift_{des}[i],\ Sift_{des}[j]\ \underset{i\neq j}{\begin{cases}\in R_{opt} & if\ \|Sift_{des}[i].x - Sift_{des}[j].x\| \le T_{opt} \\ & \&\&\ \|Sift_{des}[i].y - Sift_{des}[j].y\| \le T_{opt} \\ \notin R_{opt} & otherwise\end{cases}} \quad (1)$$

The optimized SIFT features are shown in Figure 2(b). Compared with those in Figure 2(a), the number of the optimized SIFT features have been reduced remarkably.

2.1.2 Color Features

Since SIFT features merely exploit grey information, fault detection may happen when images share similar outline while colors are quite different. To overcome this shortage, we combine the color features with SIFT to further improve the retrieval performance.

Firstly, the color features are extracted in HSV color space. *H*, *S*, *V* are divided into 8, 3, 3 bins respectively through the non-interval quantization, which are represented as *H'*, *S'*, *V'*. According to Equation (2), each one-dimensional vector consists of 72 bins. Color histogram formed by the vectors is used as the color features.

$$I=H'Q_SQ_V+S'Q_V+V' \quad (2)$$

where Q_S, Q_V are the quantization levels of *S'*, *V'*, and $Q_S=3$, $Q_V=3$.

2.2 Dimensionality Reduction

For a large-scale image database, original feature data can be further simplified by projecting the data from higher dimensional space into lower space so that it can ensure the accuracy of retrieval and reduce the computational complexity of the following processing steps as well.

LFDA is an effective dimensionality reduction method, which can reduce the dimension of the feature vector space while ensuring the recognition degree of features. In this paper, LFDA is used to reduce dimensions of features.

The projection matrix W_{lfda} consists of maximizing the local inter-class scatter matrix V_{inter} and minimizing the local intra-class scatter matrix V_{intra}. W_{lfda} can be calculated as Equation (3).

$$W_{lfda} = \arg\max_L \left| \frac{L^T V_{inter} L}{L^T V_{intra} L} \right| \quad (3)$$

where L^T is the liner transformation matrix,

$$V_{inter} = \frac{1}{2}\sum_{i,j=1}^{N} Inter_{i,j}(x_i - x_j)(x_i - x_j)^T \quad (4)$$

$$V_{intra} = \frac{1}{2}\sum_{i,j=1}^{N} Intra_{i,j}(x_i - x_j)(x_i - x_j)^T \quad (5)$$

$$Inter_{i,j} = \begin{cases} \frac{A_{i,j}}{n_c} & i,j \in C \\ 0 & when\ \ i \in C, j \notin C \end{cases} \quad (6)$$

$$Intra_{i,j} = \begin{cases} A_{i,j}\left(\frac{1}{N} - \frac{1}{n_c}\right) & i,j \in C \\ \frac{1}{N} & when\ \ i \in C, j \notin C \end{cases} \quad (7)$$

$$A_{i,j} = \frac{e^{\left(-\|x_i - x_j\|_2^2\right)}}{\rho_i \rho_j} \quad (8)$$

where N is the total number of images, C denotes a certain kind of classes, and n_c is the number of the *Cth* class. $\rho_i = \|x_i - x_i^k\|_2$, x_i^k is the k-nearest-neighbours of x_i (Zelnik-Manor L. and Perona P., 2004), and k is tuning factor. It reveals that the experimental result turns ideal when k is set as 7.

In general, the number of extracted features is much more than that of the images so that the resulting local intra-class scatter matrix V_{intra} is singular, which makes the eigenvector matrix unsolvable. In order to overcome this problem, PCA is used to reduce the dimension of the input features so that V_{intra} becomes non-singular. W_{LFDA} is calculated as in Equation (9) and (10):

$$W_{LFDA} = W_{PCA} W_{lfda} \quad (9)$$

$$W_{lfda} = \arg\max_L \left| \frac{L^T W_{PCA}^T V_{inter} W_{PCA} L}{L^T W_{PCA}^T V_{intra} W_{PCA} L} \right| \quad (10)$$

It makes the optimal effect working for both inter-class classification and intra-class.

2.3 Construction of Image Index based on Visual Vocabulary Tree

Vocabulary tree is an efficient data organization structure for image retrieval. The construction process of a vocabulary tree is shown as Figure 3.

In practice, we adopt a hierarchical *K*-means

clustering scheme to construct the vocabulary trees using SIFT features and color histogram respectively. Each feature vector is compared with the K clustering centre of each layer in a top-down manner to select the closest one until the nearest category is selected. For a vocabulary tree whose height is L, dot product operations in each layer only need K times, so the total number of calculation is KL times. Each leaf of the vocabulary tree is viewed as a visual word. The number of visual words can be calculated by Equation (11).

$$\sum_{l=1}^{L} K^l = \frac{K^{L+1} - K}{K - 1} \tag{11}$$

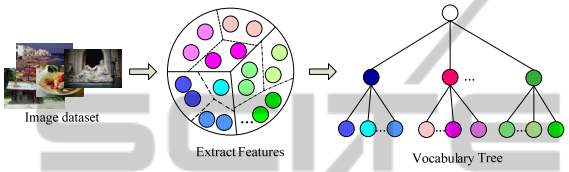

Figure 3: The process of constructing a vocabulary tree.

2.4 Similarity Measurement

The inverted index is used for large-scale image retrieval in this paper. We measure the similarity among images using the Term of Frequency-Inverse Document Frequency (TF-IDF).

In the vocabulary tree, each node i is corresponding to a visual word C_i, the term frequency of query image and database images which pass the node i are denoted as q_i and d_i, respectively. The IDF can be computed as in Equation (12):

$$IDF = \log \frac{N}{N_i} = \omega_i \tag{12}$$

where N is the total number of images, N_i is the number of images which pass the node i.

In this paper, $Q_i = q_i w_i$ denotes feature vector of query images and $D_i = d_i w_i$ denotes feature vector of images from database. The similarity between the query image and those from database can be measured by means of the L_2 norm as in Equation (13):

$$\begin{aligned} Similar(D,Q) &= \left\| \frac{Q}{\|Q\|} - \frac{D}{\|D\|} \right\|_2^2 = \sum_N |Q_i - D_i|^2 \\ &= \sum_{i|D_i=0} |Q_i|^2 + \sum_{i|Q_i=0} |D_i|^2 + \sum_{i|D_i\neq 0, Q_i\neq 0} |Q_i - D_i|^2 \\ &= 2 - 2 \sum_{i|D_i\neq 0, Q_i\neq 0} Q_i \cdot D_i \end{aligned} \tag{13}$$

By calculating the sum of products of image elements with corresponding dimension, and selecting specific ones according to similarity results, the proposed algorithm efficiently simplifies the traditional distance matching method, and thus leads to a significant improvement in retrieval speed.

In the similarity measurement, the dimension of SIFT features and color histogram are reduced and then used to construct two vocabulary trees respectively. According to Equation (14), the similarity of images is measured and ranked by comparing the index of the query image with those of the images stored in the database. The most similar top-k images will be regarded as the research results:

$$Similar = \alpha Sim_{SIFT} + \beta Sim_{HSV} \tag{14}$$

where α, β are the weights of the vocabulary trees of SIFT features and color histogram respectively. By selecting the appropriate weights, it can facilitate the image retrieval performance to reach the optimal.

3 EXPERIMENTAL RESULTS

In order to demonstrate the effectiveness of the proposed large-scale image retrieval, we performed experiments based on an image database containing 22908 colour images that are chosen from Corel database, image-searching site BaiDu, and photo-sharing site Flickr. These images are grouped in more than 50 categories, such as African people, flowers, airplane, architecture, etc.

Figure 4 shows the interface of our proposed large-scale image retrieval system. Initially, 128-dimentional SIFT features and 72-dimentional HSV color histogram features are extracted from the images. Then, LFDA is used respectively to reduce the dimension of the two features into 16. And the tuning factor k is set as 7, which will eliminate the irrelevant redundant information of the high dimensional features. Next, two vocabulary trees (branches K=10, height level=3) are constructed using hierarchical K-means clustering algorithm, which will generate 1110 visual words. The last step is similarity measurement. During this procedure, the weights of α, β are set as α=1.5, β=0.3.

The performance of image retrieval often adopts precision and recall as the evaluation criteria. Precision reflects the accuracy of a retrieval algorithm, while recall reflects the comprehensive-ness of the algorithm. The precision-recall curves of the proposed method are shown in Figure 5, where the vertical axis is precision ratio and the horizontal axis is recall ratio.

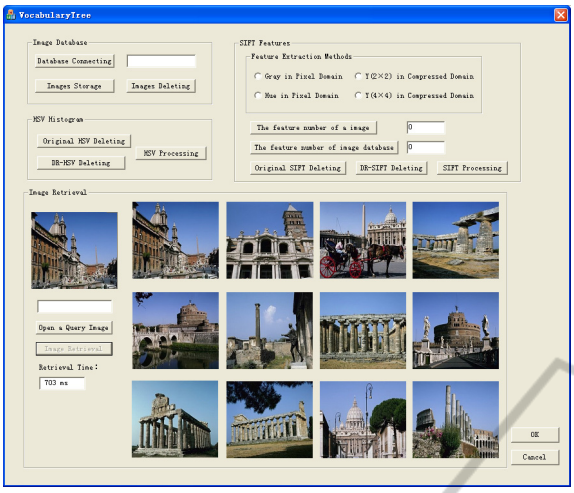

Figure 4: Large-scale image retrieval interface.

As shown in Figure 5, four curves with different colors are shown, representing the recall-precision ratio with traditional SIFT, Optimized SIFT, Optimized SIFT with HSV histogram and dimension reduced features using LFDA respectively. And Table 1 shows their precision ratio respectively.

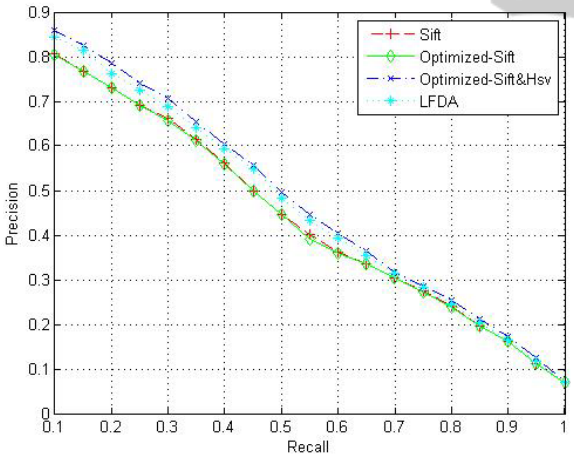

Figure 5: Precision-recall curves.

Table 1: The precision ratio using features.

	Traditional SIFT	Optimized SIFT	Optimized SIFT&HSV	LFDA
Precision ratio	80.6%	80.5%	85.9%	84.3%

From Table 1, it reveals that the performance by only using the optimized SIFT is an unsatisfactory result. However, the precision ratio of optimized SIFT is close to the traditional SIFT, which means the optimization not only maintains the performance of image retrieval, but also decreases the computational complexity of feature clustering through reducing the number of features effectively. Combining color histogram and optimized SIFT as joint image features, the precision ratio has increased obviously. Furthermore, the performance by using LFDA to reduce the dimension of feature is slightly lower than that of the one using features without dimensionality reduction.

Table 2 shows the comparison results of feature number and query index construction time of pre- and post-optimization, and similarity measurement time during the image retrieval. It can be seen that the proposed method significantly reduces the number of images features and improves the speed of query index construction.

Table 2: The comparison results of feature number and query index construction time of pre- and post-optimization, and similarity measurement time during the image retrieval.

Dataset source	Corel1K	Corel10K	Internet
Images	1,000	9,908	12,000
Feature number of pre-optimization	628,664	512,439	4,786,023
Feature number of post-optimization	309,915	356,764	2,253,502
query Index construction time of pre-optimization	380ms	290ms	380ms
query Index construction time of post-optimization	29ms	24ms	29ms
similarity measurement time	11ms	110ms	120ms

Consequently, it can be concluded that the proposed method in this paper contributes a stark reduction of computational expense through decreasing the number of SIFT features, and projecting SIFT features and color histogram from high-dimensional space to low-dimensional space, which still enable a fast and accurate image retrieval for large-scale database.

4 CONCLUSIONS

In this paper, a large-scale image retrieval based on

vocabulary tree is proposed. The method manages to reduce the number of features by optimizing SIFT descriptors and combines color histogram and optimized SIFT in order to reduce the disadvantage that traditional SIFT did not consider color information. Moreover, LFDA is adopted to reduce dimension of features while the image retrieval performance is still achieved. Finally, fast, efficient and accurate large-scale image retrieval is realized by constructing image index based on visual vocabulary tree.

ACKNOWLEDGEMENTS

The work in this paper is supported by the National Natural Science Foundation of China (No.61372149, No.61370189, No.61003289, No.61100212), the Program for New Century Excellent Talents in University (No.NCET-11-0892), the Specialized Research Fund for the Doctoral Program of Higher Education (No.20121103110017), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (No.CIT&TCD201304036), the Science and Technology Development Program of Beijing Education Committee (No.KM201410005002).

REFERENCES

Bay H., Ess A., Tuytelaars T., Van G. L., 2008. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding, Vol. 110*(3), pp. 346-359.

Böhm C., Berchtold S., Keim D. A., 2001. Searching in high-dimensional spaces - *Index structures for improving the performance of multimedia databases. ACM Computing Surveys* (CSUR), Vol. 33(3), pp.322-373.

Gionis A., Piotr I., Rajeev M., 1999. Similarity Search in High Dimensions via Hashing. In Proceedings of the *25th International Conference on Very Large Data Bases.* Morgan Kaufmann Publishers Inc. pp. 518-529.

Ke Y. and Sukthankar R., 2004. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 511-517.

Lowe D. G., 2004. *Distinctive image features from scale-invariant keypoints.* IJCV, Vol. 60(2), pp. 91-110.

Mikolajczyk K., Schmid C., 2005. A performance evaluation of local descriptors. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, pp. 1615-1630.

Nist´er D., Stew´enius H., 2006. Scalable recognition with a vocabulary tree. *In Proc. CVPR, Vol. 2*, pp. 2161-2168.

Rahulamathavan Y., Phan R. C. W., Chambers J. A., Parish D. J., 2013. Facial Expression Recognition in the Encrypted Domain Based on Local Fisher Discriminant Analysis, IEEE *Transactions on Affective Computing, Vol. 4(1),* pp.83-92.

Zobel J. and Moat A., 1998. Inverted files versus signature les for text indexing. ACM *Transactions on Database Systems, vol. 23*, pp. 453-490.

Zelnik-Manor L. and Perona P., 2004. Self-Tuning Spectral Clustering. *Proc. 18th Ann. Conf. Advances in Neural Information Processing Systems*, Vol.17, pp. 1601-1608.

http://wang.ist.psu.edu/docs/related.shtml.

http://image.baidu.com/

http://www.flickr.com/