Improving the accuracy of image search by combining the visual index and the text index

Tak-Eun Kim*, In-Hwan Hwang*, Myoung Ho Kim*
*Korea Advanced Institute of Science and Technology, Republic of Korea
E-mail : tekim, ihhwang, mhkim@dbserver.kaist.ac.kr

1. Introduction

The traditional image search engine we have used so far are based on textual descriptions attached to images. If query keywords describe the target image very well, it can return the right images that a user is interested in. Unfortunately, some images such as abstract paintings are very difficult to describe with words. For searches like that, the traditional image search engine may fail to provide the right answer because of mismatches between textual descriptions attached to images and query keywords.

Content-based image searches[1] have been emerged to overcome the problem of keyword-based image searches. It never forces users to describe an image with keywords. Instead of keywords query, it takes an image "as is" as a query. The content-based image search engine analyzes the query image, computes visual similarities between the query image and images in a database, and returns a set of images those are visually relevant to the query image. However, the search quality drops when a query image has no visual characteristics which are distinguishable from other images in a database. In this case, the images a user is really looking for are often buried under hundreds of thousands of irrelevant images.

To improve the accuracy, we propose a new approach which combines the two above approaches properly. In our approach, we utilize both an image and keywords as a query. We refer to such a query as "image-keyword joint query". Since two types of queries are considered simultaneously, an image which is relevant but having low visual similarity score to the query image now can be ranked in top result if the textual relevant score is high.

In this paper, we define a problem of image-keyword joint query processing and present a method to process the image-keyword joint query efficiently. To do that, we tightly integrate a high-dimensional visual feature index and a textual index such that both types of information can be used to prune a large amount of irrelevant search space simultaneously.

2. Related Work

To the best of our knowledge, the idea and solutions for efficient image-keyword joint query processing have not been proposed before. So there are no previous studies which can be compared directly to ours. Instead, we briefly review one research which seems to be relevant to ours. Wang et al. proposed a mobile visual search which uses multi-modal queries such as image, speech, and keywords[2]. It is much similar to our work in terms of using both image and keywords as a query. However, its goal is quite different from ours. The goal of our approach is to find images containing the same objects as the query image. However, Wang's approach is not. It is designed for finding images which are the same category as the query image.

3. Indexing Visual Features

Before discussing the proposed hybrid index structure, we discuss how to index the visual features efficiently. The visual feature is a high-dimensional vector which encodes a small patch around a salient point in a given image. Note that the visual feature is high-dimensional vector. Most widely used visual features such as SIFT exceed more than 100 dimensions. Hence, the visual features should be indexed in a high-dimensional index structure.

In order to optimize a traversal of index structure during query processing, the index adaptively changes its internal structure when data is inserted or deleted. In high-dimensional index, however, a dynamic modification of index structure produces extremely high computation overhead[3]. So, we build an index structure once and do not modify it even if data is inserted or not.

The proposed visual feature index is an m-way tree, which is built by performing hierarchical clustering of visual features extracted from training images. Starting from the root, the visual features are partitioned into m clusters. For each cluster, a clustering is performed recursively until the tree grows up to a pre-defined depth. Each non-leaf node has m reference vectors and m pointers to its child nodes. Here, a reference vector is a center vector of each cluster.

Once the tree is built, the visual features of images in a database is inserted into an appropriate leaf node of the tree. At each level of non-leaf node, a distance between the visual feature vector and each reference vector is computed and the child node which has the closest reference vector is chosen. This step performs recursively until the leaf node is chosen. Then, the image ID is assigned to the leaf node.

4. Hybrid Index Structure for Image-Keyword Joint Query Processing

The proposed index, called Image-Keyword tree or IK-tree, is a tree-based hybrid index structure which tightly integrate a visual feature index and multiple signature files. The IK-tree is built on top of the visual index we have discussed in Section 3. A signature file[4] is embedded in each tree node. A signature file is a probabilistic filter that is designed to perform a membership test whether a query item is a member of a set or not. In this paper, the signature file is used as a keyword filter. Since the signature file is a probabilistic filter, it can allow false positives. However, it never allows false negatives. That is, it returns either "possibly in a set" or "definitely not in a set".

The signature file in each node is built in bottom-up manner. Firstly, the signature file of each leaf node is built from textual descriptions of each image. Next, the signature files of non-leaf nodes are built by superimposing the signature files of its child nodes.

For a given image-keyword joint query, the search procedure traverses IK-tree from its root node and returns top-k most relevant images. Starting from the root node, the query keywords are tested on all signature files in each cluster of currently visiting node. If the membership test fails, then the sub-tree of the cluster is pruned. If the membership test passes, then a visual relevance score, which is a Euclidean distance between a visual feature of the query image and a reference vector of the cluster, is computed. The visual relevance score is used for final re-ranking of candidate images. Since most search spaces are pruned earlier, we can get the result quickly.

5. Experiments

To measure the performance of IK-tree, we designed a baseline algorithm which filters irrelevant images by scanning whole images in a database one-by-one. During scanning of images, an image is simply filtered if keyword tags of the image does not contain the query keyword. For candidate images which are not filtered during scanning step, visual similarity scores are calculated and top-k images are returned as a final result.

All the algorithms were written in MATLAB, and all experiments were performed on Intel i5 Linux machine equipped with 8GB of RAM and 128GB SSD drive. The experiments were performed using mirflickr-1M database, which is a collection of one million images crawled from Flickr. The number of unique keywords on the mirflickr-1M dataset is approximately 815,000. Each image has 50 keyword tags on average. For simplicity, we set the length of each signature file as 8,192 bits (or 1 kilibytes).

We measured query response times with respect to the number of query keywords. The query response time, on average, were 30, 110 and 170 milliseconds for 1, 2 and 3 keywords, respectively. The baseline approach takes more than 3,000 milliseconds on average. Although the baseline approach is not enough to prove the superiority of the IK-tree, such query response time (30 to 170 milliseconds) is considered as a "good enough" in other indexes.

6. Conclusion and Future Work

In this paper, we defined a new kind of query, called image-keyword joint query, and proposed an index structure, called IK-tree, for efficient processing of joint queries. Our solution integrates a tree-based visual index and signature files tightly so that the integrated index structure can filter both visual and textual information simultaneously.

There are a number of open problems that we have to address. The weak point of IK-tree is that the pruning power of signature files drop rapidly when superimposition of signature files performed multiple times[5]. Since signature files are built in bottom-up manner, most signature files in lower-level nodes have almost zero pruning powers. To solve such problem, we can consider varying size of signature files, where the size depends on the level of node.

7. Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning) of Korea under the ITRC support program(NIPA-2013-H0301-4009), and the National Research Foundation of Korea grant funded by the Korea government(MEST) (No.2012R1A2A2A01046694).

8. References

[1] J. Sivic, and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos", In Proc. of ICCV, Nice, France, October 13-16, 2003.

[2] Y. Wang, T. Mei, J. Wang, H. Li, and S. Li, "JIGSAW: interactive mobile visual search with multimodal queries", In Proc. of MM, Scottsdale, Arizona, USA, November 28-December 1, 2011.

[3] D. Nister, and H. Stewenius, "Scalable Recognition with a Vocubulary Tree", In Proc. of CVPR, New York, USA, June 17-22, 2006.

[4] D. L. Lee, Y. M. Kim, and G. Patel, "Efficient signature file methods for text retrieval", TKDE, Vol. 7, Issue 3, pp. 423-435, 1995.

[5] I. D. Felipe, V. Hristidis, and N. Rishe, "Keyword Search on Spatial Databases", In Proc. ICDE, Cancun, Mexico, April 7-14, 2008.