Content Based Image Retrieval on the Web

David Novak, Michal Batko and Pavel Zezula

Masaryk University, Brno, Czech Republic
{david.novak|batko|zezula}@fi.muni.cz

It has become customary that practically any information can be in digital form. Searching through future internet will be complicated because of: (1) the diversity of ways in which specific data can be sorted, compared, related, or classified, and (2) the exponentially increasing amount of digital data. Accordingly, a successful search engine should address problems of *extensibility* and *scalability*.

We present and demonstrate capabilities of MUFIN (Multi-Feature Indexing Network)[1]. To achieve an independence of the similarity abstraction and the ability to process large collections of data, MUFIN is based on two basic paradigms: (1) the *metric space* model of similarity, and (2) the concept of *structured Peer-to-Peer* (P2P) networks. The metric space model of similarity has already proved to be a very powerful concept for expressing many different forms of similarity of vectors, strings, sets and other data types. Most of the available technologies for processing metric data as well as important foundations of such approach are described in the recent book [1].

At the same time, P2P structures have emerged as an attractive architectural paradigm for the management of large data collections. In general, P2P networks are distributed and fully-decentralized systems that aim at resource sharing and parallel processing. They constitute a promising building block of next-generation search engines that are expected to deal with huge amounts of heterogeneous (e.g. textual, multimedia, scientific, etc.) and continuously changing data. Though available structured P2P methods are prevalently designed to support execution of simple keyword queries – these methods are also known as the *distribute hash tables* (DHT) – they have also been applied to the metric searching with great success.

Considering both the extensibility and scalability objectives, MUFIN is quite unique and thus difficult to compare with other previously proposed systems. Its capabilities are achieved by exploiting and extending the research achievements in the area of metric searching technology over the last decade. We illustrate the versatility of this approach on the content-based image and face recognition search engines. The scalability is demonstrated by an interactive image retrieval system that indexes 50 million images. Compared to other existing search systems (e.g. Alipr[2] or TinEye[3]), MUFIN indexes an image database of one or two orders of magnitude bigger. Moreover, these systems are restricted to vector-space models of image features, and, unlike MUFIN, cannot be efficiently used with different measures of similarity or applied on different datasets. Further-

[1] http://mufin.fi.muni.cz/
[2] http://alipr.com/
[3] http://tineye.com/

more, the scalability grows with constant complexity, which is very important for future internet technologies. Thanks to a recent IBM SUR (Shared University Research) Award granted to our university, a stable demo will soon be available as a web service for even larger data collections.

Image Retrieval Demo

We will demonstrate capabilities of MUFIN through content-based similarity search in a collection of 50 million digital images. The indexing and searching engine is formed by a combination of a distributed system M-Chord [2, 3] and a centralized indexing structure M-Tree [4]. This approach is very general in various aspects, as it is based purely on the metric data model [1] and the M-Chord follows the paradigm of structured peer-to-peer networks.

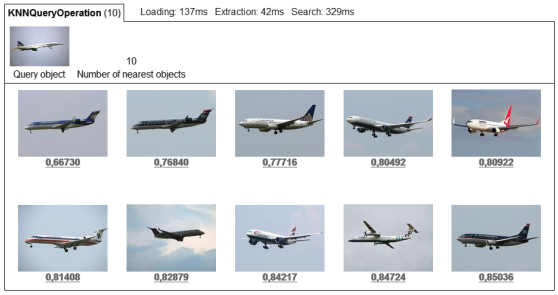

Fig. 1. Example of a similarity query result.

Demo web page: `http://mufin.fi.muni.cz/search/`
An online preliminary demonstration of the image search. Please note that the site might be temporarily unavailable from time to time.

Characteristics:

- The dataset consists of 50,000,000 images from a photo-sharing system Flickr[4] and five MPEG-7 features extracted from each image (see below). The data forms a part of the CoPhIR database[5] created within EU project SAPIR[6].
- Following the query-by-example concept, the system retrieves k images which are the most similar to the query image.
- The visual similarity is measured by five MPEG-7 descriptors [5]; their dissimilarity is measured by the following metric functions and they are aggregated into a single metric function using the following weights:

[4] http://www.flickr.com
[5] Content-based Photo Image Retrieval Test-Collection, http://cophir.isti.cnr.it
[6] EU IST FP6 project 045128 (SAPIR), http://www.sapir.eu

MPEG-7 Feature	Metric	Weight
Scalable Color	L_1 metric	2
Color Structure	L_1 metric	3
Color Layout	sum of L_2 metrics	2
Edge Histogram	special [5]	4
Homogeneous Texture	special [5]	0.5

- The query image can be either taken from the indexed database or it can be any image on the Web – see an example of the query result in Figure 1.
- The distributed indexing and searching system is composed of 2000 peers operated on a hardware infrastructure with 32 CPUs; the Web-based client connects remotely to any of the peers.
- The response times of the system are about one second and the query throughput is about ten queries per second.

Face Recognition Demo

To demonstrate the generality (extensibility) of the metric-based MUFIN approach we will show another prototype application: a face-recognition and retrieval system.

Characteristics:

- The data we have used represent faces detected in photos obtained from a digitized photo gallery of the Masaryk University. The photo gallery contains about 10,000 photos on which about 16,000 faces were detected.
- For each detected face a facial descriptor was acquired using the method described in the MPEG-7 standard as the Advanced Face Descriptor (AFR) [5].
- The similarity of two faces is measured by applying a weighted L_1 metric on the AFR, which is actually a vector of 96 dimensions.
- The extracted face descriptors are organized in a local M-Tree [4] response times below a half a second.

References

1. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: The Metric Space Approach. Volume 32 of Advances in Database Systems. Springer-Verlag (2006)
2. Novak, D., Zezula, P.: M-Chord: A scalable distributed similarity search structure. In: Proceedings of INFOSCALE '06, Hong Kong, ACM Press (2006)
3. Batko, M., Novak, D., Falchi, F., Zezula, P.: On scalability of the similarity search in the world of peers. In: Proceedings of INFOSCALE 2006, Hong Kong, New York, ACM Press (2006) 1–12
4. Ciaccia, P., Patella, M., Zezula, P.: M-Tree: An efficient access method for similarity search in metric spaces. In: Proceedings of VLDB'97, August 25–29, 1997, Athens, Greece. (1997) 426–435
5. Manjunath, B., Salembier, P., Sikora, T., eds.: Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley & Sons, Inc., New York, NY, USA (2002)