1 Introduction

Similarity is a central notion throughout human lives. In perception, the similarity between sets of visual or auditory stimuli influences the way in which they are grouped. In speech recognition, the similarity between different phonemes determines how confusible they are. In classification, the category of a new instance may be influenced by the similarity of the new instance to past instances or a stored prototype. In memory, it has been suggested that retrieval of a cue depends on the similarity of past memory traces to the representation of the cue. Since almost everything that we see, read, hear, write and measure is, or very soon will be, available in digital forms, computer systems must support similarity. But the growth of digital material is posing another big challenge: The exponential increase of data volume makes the scalability a matter of concern. We believe that the core ability of future data processing systems is the similarity management of very large and growing volumes of data.

2 MUFIN Architecture

The four-tier architecture of the MUFIN system is schematically depicted in Figure 2. The executive core of the system is formed by several independent P2P overlays in the first tier. Each overlay maintains only the data of a specific feature, e.g. shape descriptors of images, color histograms, or protein spectra vectors. While operating, overlay’s peers exchange messages, hence they are completely independent of the specific hardware infrastructure. In the second tier, the independent overlays are linked together into a single virtual multi-overlay system where any peer from any overlay can access all the other overlays. For instance, a peer maintaining shapes of images can ask for similar color histograms. Interfaces for maintaining data and posing queries represent the third tier of the architecture. In particular, MUFIN offers a native API, a web service interface and a plug-in interface for linking with external indexes. Finally, the fourth level comprises of a variety of user interfaces.

MUFIN is a complex Java software built by means of the Metric Similarity Search Implementation Framework MESSIF [1], which is a Java library of implementation tools giving a flexibility of applying suitable search strategies for specific purposes.

The properties of MUFIN can be characterized by the following features.
Extensibility Different similarity search indexes for specific applications can be built with a single tool.

Scalability The underlying P2P technology allows the system to adapt to extremely large datasets without sacrificing performance.

Multi-layer system Since similarity on a specific collection of objects can be seen from different similarity points of view, multiple overlays can be built.

Complex query processing In order to adjust effectiveness of search according to needs of individual users, several overlays can be combined together. The system can also utilize external services that support a sorted access.

Approximation and performance tuning In order to further improve performance, approximation techniques can be applied. The level of approximation can be adjusted manually by users as well as automatically by the system according to the current load.

Infrastructure independence Instances of MUFIN can operate on a wide variety of hardware infrastructures ranging from a local network of common workstations to a world-wide GRID system.

3 MUFIN: Similarity Image Search

We demonstrate an “instance of MUFIN” designed for content-based search on large databases of general digital images. The dataset consists of 100 million images taken from CoPhIR Database\(^1\). Each image is represented by five global MPEG-7 descriptors [3] aggregated into a single metric space and the system retrieves \( k \) images which are the most similar to a given query image (according to the aggregated metric).

![Example of similarity query result.](http://mufin.fi.muni.cz/imgsearch/)

The data is indexed by a P2P-based data structure M-Chord [4] with 2000 logical peers. Peers organize their data locally in an M-Tree [2]. The system physically runs on six IBM servers (two quad-core CPUs, 16G RAM, six disks with RAID 5). The search is demonstrated via a Web-based interface that is available online at [http://mufin.fi.muni.cz/imgsearch/](http://mufin.fi.muni.cz/imgsearch/). Figure 3 shows an example of a MUFIN query result.

---

1\(^{\text{CoPhIR: Content-based Photo Image Retrieval Database: http://cophir.isti.cnr.it/}}\)

4 Current CBIR Systems

Several systems for Content-based Image Retrieval (CBIR) are currently available:

**ALIPR** searches a set of images according to automatically generated annotations [http://www.alipr.com],

**ImBrowse** allows to search about 750,000 images by color, texture, shapes (and combinations) employing five independent engines [http://media-vibrance.itn.liu.se],

**idée** searches a commercial database of 2.8 million images according to image signatures [http://labs.ideeinc.com],

**GazoPa** is a private service by Hitachi searching 50 million images by color and shape [http://www.gazopa.com].

All current CBIR systems, except for the very recent project GazoPa, have databases two orders of magnitude smaller than MUFIN. Moreover, approaches based on signatures usually work well only for near-duplicates. The mentioned systems are designed only for searching digital images by a specific method, which is in contrast with the highly versatile approach of MUFIN.

5 Conclusions

By exploiting and extending the research achievements in metric searching technology over the last ten years, MUFIN proves the expected extensibility and scalability properties of this technology. The scalability has been demonstrated by an interactive image retrieval system over 280-dimensional vectors, which is one order of magnitude higher than what most of the literature considers to be the dimensionality curse. The system organizes a 100-million image collection, which is two orders of magnitude more than what similar systems use, to the best of our knowledge.

References


