Enabling Content-Based Image Retrieval in Very Large Digital Libraries

Paolo Bolettieri, Andrea Esuli, Fabrizio Falchi, Claudio Lucchese, Raffaele Perego, and Fausto Rabitti

ISTI-CNR, Pisa, Italy.
firstname.lastname@isti.cnr.it

Abstract. Enabling effective and efficient Content-Based Image Retrieval (CBIR) on Very Large Digital Libraries (VLDLs), is today an important research issue. While there exist well-known approaches for information retrieval on textual content for VLDLs, the research for an effective CBIR method that is also able to scale to very large collections is still open. A practical effect of this situation is that most of the image retrieval services currently available for VLDLs are based only on textual metadata. In this paper, we report on our experience in creating a collection of 106 million images, i.e., the CoPhIR collection, the largest currently available to the scientific community for research purposes.We discuss the various issues arising from working with a such large collection and dealing with a complex retrieval model on information-rich features. We present the non-trivial process of image crawling and descriptive feature extraction, using the European EGEE computer GRID. The feature extraction phase is often ignored when discussing the scalability issue while, as we show in this work, it could be one of the toughest issues to be solved in order to make CBIR feasible on VLDLs.

1 Introduction

Everybody knows about the data explosion. According to recent studies, in the next three years, we will create more data than has been produced in all of human history. Regarding images, the Enterprise Strategy Group[1] estimates that more than 80 billion photographs are taken each year. Storing them would require 400 petabytes of storage. Therefore the management of digital images promises to emerge as a major issue in many areas providing a lot of opportunities in the next years, particularly since a large portion of pictures still remains as "unstructured data", i.e., with no meaningful associated tags.

Current searching engines headed by Google are in the center of current information age; Google answers daily more than 200 million queries against over 30 billion items. However, the search power of these engines is typically limited to text and its similarity. Since less than 1% of the Web data is in textual form, the rest being of multimedia/streaming nature, we need to extend our next-generation search to accommodate these heterogeneous media.

[1] http://www.enterprisestrategygroup.com/

A typical approach adopted by the current engines to provide search functionality on these data types is to leverage on the textual information that is potentially associated with the non-textual content, e.g., the text nearby the image published in a Web page, or other attributes, e.g., metadata fields reporting the name of the author, the title, or a list of tags. This approach is adopted by almost all the Web scale engine (e.g., Google, Yahoo, Ask) but also can be found in the latest VLDL initiatives, such as Europeana[2].

An orthogonal approach is the Content-based Image Retrieval (CBIR). It is not a new area as demonstrated by a recent survey [1] which reports on nearly 300 systems, most of them exemplified by prototype implementations. However, the typical database size is in the order of thousands of images. Very recent publicly-available systems, such as ImBrowse, Tiltomo or Alipr[3], declare to index hundreds of thousands of images. There is a huge discrepancy between these numbers and the volumes of images available on current Web, so we decided to investigate the situation by shifting the current bounds up by two orders of magnitude. The result of this process is the CoPhIR (**Co**ntent-based **Ph**oto **I**mage **R**etrieval) collection [2], a collection of 106 million images crawled from the Flickr[4] photo sharing website. The size of the CoPhIR collection goes very far beyond the current practice.

Each image has associated a large number of metadata fields, from the title, description and comments, to the date and the geographical coordinates of the location where the image has been acquired. The availability of a rich amount of metadata information is a nice feature of the collection, that allows to use it to build information-intensive search systems, which is a typical situation in digital libraries. We have enriched the information associated to each image by extracting five MPEG-7 visual features, each one describing some specific characteristic of the image content.

As we detail in Section 2 the image crawling and feature extraction process resulted to be computationally very expensive: the entire process would have required about 12 years on a standard PC. We have used the European EGEE computer GRID to reduce this processing time to just a few days, showing how distributed computing is a crucial element to enable CBIR on the large scale. The CoPhIR collection has been made freely available to the scientific community, and it has been already used to develop and test many CBIR systems.

A relevant result of this experience is that we have observed how the feature extraction process, which is often ignored when discussing the scalability issues of CBIR systems, is probably the toughest issue to be solved in order make CBIR, and in general content-based retrieval for non-textual data, feasible on VLDLs.

---

[2] http://www.europeana.eu/

[3] http://media-vibrance.itn.liu.se/, http://www.tiltomo.com/, http://www.alipr.com/

[4] http://www.flickr.com/

2 Building the Image Collection

Collecting a large amount of images for investigating CBIR issues is not an easy task, at least from a technological point of view. The challenge is mainly related to the size of the collection we are interested in. Shifting the current state-of-the-art bounds of two orders of magnitude means building a 100 million images collection, and this size makes very complex to manage every practical aspect of the gathering process.

2.1 Choosing the Data Source

Given our need of high-quality data, we decided to crawl one of the popular photo sharing sites born in the last years with the goal of providing permanent and centralized access to user-provided photos. This approach has several advantages:
Image Quality. Photo sharing sites mainly store high-quality photographic images.
Collection Stability. These sites provide quite static, long term and reliable image repositories.
Legal Issues Storing for a long time a publicly available image may in some case violate author's copyrights. Since Photo sharing sites are fairly static, we can build a quite stable collection without storing the original files, but maintaining only the hyperlinks to the original photos.
Rich Metadata Photo sharing sites provide a significant amount of additional metadata about the photos hosted, as described in Section 2.2. The availability of rich metadata is a relevant aspect with respect to experimentation, because it allow to investigate on new models of combination of traditional, metadata-based search with content-based search.

Among the most popular photo sharing sites, we chose to crawl Flickr, since it is one with the richest additional metadata and provides an efficient API[5] to access its content at various levels.

2.2 Crawling the Flickr Contents

In February 2007, we crawled the graph of Flickr users, selecting about one million of users. We then exploited the Flickr API to get the whole list of public photos owned by each of these users, retrieving 300 million distinct photo IDs.

For each photo we decided to retrieve almost all information available: title and description, identification and location of the author, user-provided tags, comments of other users, GPS coordinates, notes related to portions of the photo, number of times it was viewed, number of users who added the photo to their favorites, upload date, and all the information stored in the EXIF header of the image file. In order to support content based search, we extracted several MPEG-7 *visual descriptors* (VDs) from each image [3]. A VD characterizes a particular visual aspect of the image. They can be, therefore, used to identify images which have a similar appearance. VDs are represented as vectors, and

[5] http://www.flickr.com/services/api/

the MPEG-7 group proposed a distance measure for each descriptor to evaluate the similarity of two objects [4]. We have chosen five MPEG-7 VDs: Scalable Color, Color Structure, Color Layout, Edge Histogram, Homogeneous Texture.

The extraction of MPEG-7 visual descriptors from high-quality images is computationally very expensive. The MPEG-7 eXperimentation Model (MPEG-7 XM) software running on a AMD Athlon XP 2000+ box takes about 4 seconds to extract the five features from an image of size 500×333 pixels. We can estimate that a single standard PC would need about 12 years to process a collection of 100 million images. It was thus clear that we needed a large number of machines working in parallel to achieve our target collection of 100 million images in a reasonable amount of time.

We developed an application that allows to process images in parallel on an arbitrary (and dynamic) set of machines. This application is composed of three main components: the *image-id server*, the *crawling agents*, and the *repository manager*. The image-id server provide crawling agents with a number of photo identifiers to be processed. The crawling agent asks the image-id server for a set of image identifiers to process. The agent downloads a medium resolution version of the image (500 pixel on the larger side) and all the metadata information associated to it. Then the MPEG-7 XM software is used to extract the aforementioned five VDs. The extracted features and all the available metadata are used to produce an XML file containing the knowledge about the image. In fact, each entry of the CoPhIR collection is an XML structure containing: (i) identification information that allows to link and retrieve the corresponding image on the Flickr Web site; (ii) the image textual data and metadata: author, title, description, GPS location, tags, comments, view count, etc.; (iii) an XML sub-structure containing the information related to the five MPEG-7 visual descriptors.

Given that crawling agent is the most computational-demanding component of our application, we have considered GRID to be the right technology to obtain large amount of computing power we needed, as described in Section 2.3.

The repository manager runs on a large file-server machine providing 10TB of RAID storage. It receives and stores the processed data by the crawling agents.

2.3 Using the GRID for Crawling and Feature Extraction

GRID is a very dynamic environment that allows to transparently run a given application on a large set of machines. In particular, we had the possibility to access the EGEE (Enabling Grids for E-sciencE) European GRID infrastructure[6] provided to us by the DILIGENT (Digital Library Infrastructure on Grid Enabled Technology) IST project[7]. We were allowed to use 35 machines spread across Europe. We did not have an exclusive access to these machines and they were not available all the time. Both hardware and software configurations were heterogeneous.

[6] http://www.eu-egee.org/
[7] http://www.diligentproject.org/

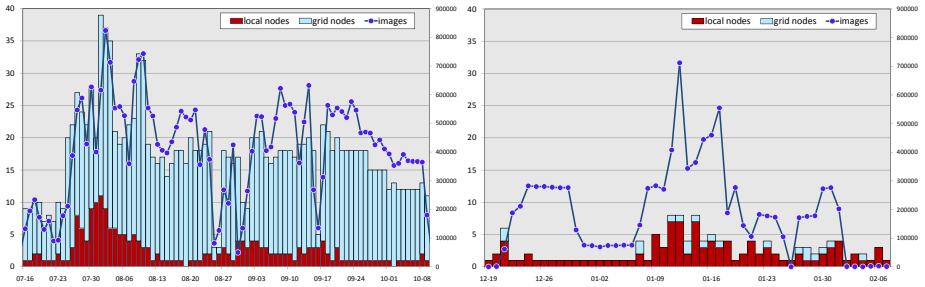

Fig. 1. Number of GRID and local machines available during the two crawling periods: from July 16^{th} to October 9^{th} 2007 (left) and from December 19^{th} 2007 to February 7^{th} 2008 (right).

The crawling agent is logically divided into two modules. The first one manages the communication with the image-id server,uses Flickr APIs, and sends the results to the repository manager. Due to the latencies of the crawling task, the crawling agent can instantiate a number of threads, each of them taking care of processing a different image. The settings which proved well is to have four threads per agent and to process a set of 1,000 images. These parameters induced computations times of 20 to 60 minutes depending on the CPU speed. The second module of the crawling agent is the feature extraction component based MPEG-7 XM software.

Submitting a job to a GRID infrastructure, the user does not have a full control on the time and location where the job runs. The GRID middleware software accepts the job description and schedules it on the next available machine according to internal policies related to the load of each node, the priority of the different organization using the GRID infrastructure, etc. The GRID provides a best-effort service, meaning that a job submitted to the GRID may be rejected and never executed. Indeed, there are several factors that may cause the failure of a job submission. Out of the 66,440 jobs submitted, only 44,333 were successfully executed that means that 33,3 % of the jobs failed for GRID resources unavailability.

Our straightforward approach together with the self-scheduling of images by each crawling agent has two important advantages. First, in case the GRID middleware is not able to deploy the given job, there would be no consequences in the remainder of the system, especially, no image will be skipped. Second, in case of a software update, it is just needed to replace the old version on the repository manager with the new one.

The crawling process took place in two separate periods, both because of GRID availability and because we needed to consolidate the data after the first period. In Figure 1, we report on the number of machines available during the crawling process. During the first period, the GRID provided an average of 14.7 machines out of the 35 and, simultaneously, there were 2.5 local machines available, on average. Also the availability of the machines during the day was unstable: the local machines were mainly available over night while some of the GRID machines were available only for a few hours per day.

2.4 The CoPhIR Collection

The result of the complex crawling and image processing activity described above is the CoPhIR collection. The data collected so far represents the world largest multimedia metadata collection available for research purposes, containing visual and textual information regarding 106 millions images. Given the effort required in building a such large test collection, and the potential interest to the international research community, in order to make experiments in large-scale CBIR, we decided to make it available outside the SAPIR project scope.

The disk space requirement for the CoPhIR collection consist of 245.3 GB for the XML data, 54.14 GB for the image content index, and 355.5 GB for the image thumbnails. CoPhIR images come from 408,889 distinct authors, with a top contributor of 156,344 images (user *conrado4*), and a *median* value of images per author equal to 69. The total number of comments in the collection is 55,188,775. The total number of tag instances is 334,254,683, from a set of 4,666,256 distinct tags. Each image is thus associated on average with 0.52 comments and 5.02 tags.

In the collection, 66,532,213 images (62.77% of the whole CoPhIR) have *popularity* information, i.e., number of views and number of selections as *favorite*. The average number of views per image is 41.7, with a top value of 599,584 views. Just half of the images with popularity information have registered more than 2 views (32,723,369 images, 49.18%), and only 4,963,257 images (7.46% of the part of CoPhIR with popularity information) have been marked has favorite by at least one Flickr user. 8,655,289 images (8.17% of the whole CoPhIR) have *geolocation* information associated to them.

3 Cases of use of the CoPhIR collection

In this section we report some information about research projects and indexing techniques which were able to index the whole CoPhIR collection (i.e., 106 M images) for content-based searching. All of them are using a single metric function defined as a weighted sum of the individual feature distances suggested by the MPEG group. The weights have been reported in [5] and have been defined following the studios reported in [6, 7].

3.1 SAPIR

The scalability challenge is the focus of the European project SAPIR (Search on Audio-visual content using Peer-to-peer Information Retrieval)[8] that aims at finding new ways to analyze, index, and retrieve the tremendous amounts of speech, image, video, and music that are filling our digital universe. CoPhIR has been built and used in SAPIR for experimenting various approach: text based, similarity search, geographic search, etc. All P2P index and search techniques [8] are based on the metric space model and implemented over the same framework - the Metric Similarity Search Implementation Framework (MESSIF) [9].

[8] SAPIR European Project, IST FP6: http://www.sapir.eu/

For the scope of improving throughput and response time, during the SAPIR project a metric cache was developed [10]. Unlike traditional caching systems, the proposed a caching system might return a result set also when the submitted query object was never seen in the past. In fact, the metric distance between the current and the cached objects is used to drive cache lookup, and to return a set of approximate results when some guarantee on their quality can be given.

3.2 Metric Inverted File

Amato and Savino [11] have proposed the *Metric Inverted File* (MIF), a novel method for approximate similarity search based on the use of *permutations* to represent the indexed objects. A permutation, assigned to each indexed object, consists of a list of the elements of a set of *reference object*, sorted by their order of similarity with the object. The MIF builds an inverted-list-based data structure on permutations, which allows to approximate the exact order of similarity of the indexed objects, with respect to a query and the similarity measure in use, by efficiently computing a similarity measure on permutations. Although the authors report in [11] results of experiments on a smaller collection, a demo of MIF on the CoPhIR collection is available[9], showing a good trade-off on efficiency/effectiveness in the retrieval process.

3.3 MiPai

MiPai[10] [12] is a CBIR system based on the PP-Index [13] data structure for approximated similarity search. The PP-Index is an efficiency-aimed data structure that belongs to the family of the permutation-based indexes [11]. It has nice parallelization properties that allow to tune the efficiency/effectiveness trade-off with respect to the available hardware resources. Experiments on the CoPhIR collection have shown that the PP-Index is able to achieve almost exact results (i.e., 97% precision with respect to the exact order defined on the similarity measure in use) within just a few seconds, while an exaustive search process would take more than an hour. Moreover, it is interesting to note that it took just 12.5 hours to build a PP-Index on the whole CoPhIR collection on a single PC, showing how the phase of index creation from the extracted feature can be orders of magnitude smaller than the preceding phase of features extraction.

4 Conclusions

No doubts that the scalability issue for new digital data types is a real issue, which can be nicely illustrated by difficulties with the management of the fast growing digital image collections. In this paper, we focus on a strictly related challenge of scalability: to obtain a non-trivial collection of images with the corresponding descriptive features.

We have crawled a collection of over 100 million high-quality digital images, which is almost two orders of magnitude larger in size than existing image

[9] http://mi-file.isti.cnr.it/CophirSearch/
[10] http://mipai.esuli.it/

databases used for content-base retrieval and analysis. Using a GRID technology, we have extracted five descriptive features for each image. This information is kept handy in XML files – one for each image – together with the metadata and links to original images in Flicker. This unique collection is open to the research community for experiments and comparisons. More than 50 research institution worldwide already asked access to the CoPhIR collection by registering at the CoPhIR Web site[11], and by signing the CoPhIR Access Agreement, which establishes conditions and terms of use for the collection.

References

1. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (2008) To appear.
2. Bolettieri, P., Esuli, A., Falchi, F., Lucchese, C., Perego, R., Piccioli, T., Rabitti, F.: CoPhIR: a test collection for content-based image retrieval. CoRR abs/0905.4627 (2009)
3. ISO/IEC: Information technology - Multimedia content description interfaces. (2002) 15938.
4. Manjunath, B., Salembier, P., Sikora, T., eds.: Introduction to MPEG-7: Multimedia Content Description Interface. J. Wiley & Sons, New York, USA (2002)
5. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubsky, J., Zezula, P.: Building a web-scale image similarity search system. Multimedia Tools and Application (2009) forthcoming.
6. Amato, G., Falchi, F., Gennaro, C., Rabitti, F., Savino, P., Stanchev, P.: Improving image similarity search effectiveness in a multimedia content management system. In: Proc. of Workshop on Multimedia Information System (MIS). (2004) 139–146
7. Stanchev, P., Amato, G., Falchi, F., Gennaro, C., Rabitti, F., Savino, P.: Selection of mpeg-7 image features for improving image similarity search on specific data sets. In: Proceedings of the 7-th IASTED International Conference on Computer Graphics and Imaging (CGIM 2004), ACTA Press (August 2004) 395–400
8. Batko, M., Novak, D., Falchi, F., Zezula, P.: Scalability comparison of peer-to-peer similarity search structures. Future Generation Computer Systems **24**(8) (2008) 834 – 848
9. Batko, M., Novak, D., Zezula, P.: MESSIF: Metric similarity search implementation framework. In: Proc. of DELOS Conference. Volume 4877 of LNCS. (2007) 1–10
10. Falchi, F., Lucchese, C., Orlando, S., Perego, R., Rabitti, F.: Caching content-based queries for robust and efficient image retrieval. In: EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology, ACM (2009) 780–790
11. Amato, G., Savino, P.: Approximate similarity search in metric spaces using inverted files. In: INFOSCALE '08: Proceeding of the 3rd International ICST Conference on Scalable Information Systems, Vico Equense, Italy (2008) 1–10
12. Esuli, A.: MiPai: using the PP-Index to build an efficient and scalable similarity search system. In: SISAP '09, Proceedings of the 2nd International Workshop on Similarity Search and Applications, Prague, CZ (2009) 146–148
13. Esuli, A.: PP-index: Using permutation prefixes for efficient and scalable approximate similarity search. In: 7th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR'09). (2009) 17–24

[11] http://cophir.isti.cnr.it