CoPhIR

The CoPhIR collection

The data collected so far represents the world largest multimedia metadata collection available for research purposes, with the target to reach 100 million images within the end of 2008.

- 54,585,718 images
- 355.5 GB of thumbnails
- 245.3 GB of XML description files

An XML description file stores five MPEG-7 descriptors (scalable colour, colour structure, colour layout, edge histogram, homogeneous texture) and other textual data. On average each photo is annotated with 3.1 tags, has been viewed 42 times, and received 0.53 user comments.

The CoPhIR collection is available outside the SAPIR project scope for scientific research. The collection complies to the most restrictive Creative Commons license, to the European Recommendation 29-2001 CE, based on WIPO (World Intellectual Property Organization) Copyright Treaty and Performances and Phonograms Treaty, and to the current Italian law 68-2003.

Contacts

http://cophir.isti.cnr.it
cophir@isti.cnr.it

Credits

Paolo Bolettieri
Fabrizio Falchi
Claudio Gennaro
Claudio Lucchese
Matteo Mordacchini
Raffaele Perego
Tommaso Piccioli
Fausto Rabitti

High Performance Computing
and
Networked Multimedia Information Systems
laboratories at

CONSIGLIO NAZIONALE DELLE RICERCHE

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

SAPIR
Search in Audio-Visual
http://www.sapir.eu

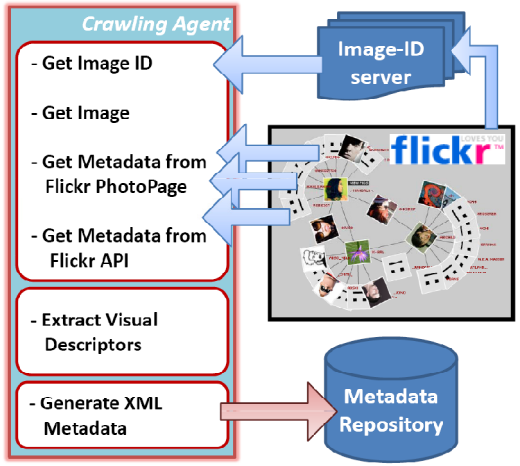
CoPhIR

Content-based Photo Image Retrieval Test-Collection

http://www.flickr.com/photos/auntiep/35704921

The largest multimedia metadata collection

CoPhIR Content-based Photo Image Test-Collection



Mission

During the last decade, we witnessed an increasing deal of interest in Content-Based Image Retrieval (CBIR). Nevertheless, the absence of a publicly available significantly large multimedia collection is still hindering academic research.

Our goal is to provide the scientific community with a collection of 100 millions metadata-rich high quality pictures for content-based image retrieval.

Building the collection

The CoPhIR collection is based on Flickr™ images and related metadata. We selected a subset of Flickr™ users and accomplished a crawling process that collected their public photos. For each photo, the CoPhIR collection contains an XML description of its salient features to be used by a retrieval system.

For each photo, the CoPhIR collection stores its title, description, author, tags, comments, notes, and also its GPS coordinates, the number of views and the number of users considering the photo a favourite.

More importantly, the CoPhIR collection stores five MPEG-7 visual descriptors of every image. These are vital for any content-based similarity search system. Scalable colour, colour structure, colour layout, edge histogram and homogeneous texture descriptors, compliant with the MPEG-7 standard, were extracted from a mid-resolution version of each photo, having width of 500 pixels.

The cost of extracting such features is tremendous. Indeed, it would take 6 years to process 50 millions images on a single machine. We thus used a large number of machines made available by the EGEE grid to build the CoPhIR collection in about three months.

73 distinct machines spread across Europe were used to extract the MPEG-7 visual descriptors.

The result of this complex crawling and image processing activity is a huge and multifaceted multimedia collection. A wide spectrum of techniques based either on the text or on the social context associated with an image may be used to experiment the effectiveness of a retrieval system.

The CoPhIR collection has served as the basis of the experimentations of content-based image retrieval techniques and their scalability, in the context SAPIR, an IST FP6 project for Search in Audio Visual Content Using Peer-to-peer IR.