Content-Based Search in Distributed Multimedia Databases: Addressing Heterogeneity and Provider-Side Privacy

François Boisson, Michel Crucianu, Dan Vodislav
Vertigo - CEDRIC & Wisdom, CNAM, 292 rue St. Martin, 75141 Paris cedex 03, France
{Francois.Boisson, Michel.Crucianu, Dan.Vodislav}(a)cnam.fr

The evolution of user requirements and of enabling technologies will have a significant impact on how online search for multimedia content is performed. A major demand of the users is to avoid the "keyword bottleneck", in part by the inclusion of *content-based* search criteria. But indexing the content of images, videos or music is very different from indexing hypertext, both because of the nature and volume of the data and because of specific rights issues. To highlight some research challenges that we deem important in this context, we start by taking a closer look at two major types of existing or potential multimedia content providers: the general public and large institutional archives.

With widespread digital imaging and cheap high-capacity storage, end users became creators and potential providers of multimedia content. According to the prevailing paradigm for access to user-generated content (UGC), producers upload their content to the central servers of a provider that makes it available to a large public. Consequently, content creators have difficulties in keeping control over their content. This approach also has scalability problems because of the reliance on central servers. One solution could consist in the use by the provider of a distributed system instead of central servers and in the setting up of comprehensive contractual relations between the creators and the provider. Another solution, that is likely to coexist with the previous one, would be the establishment of a peer-to-peer (P2P) network where every content producer is also a provider. To make search possible in such a network, a global index has to be created and then distributed among the peers. The recent proposals that focus on content-based search for multimedia and follow this paradigm assume that all the peers employ the same content descriptors and agree with their dissemination over the network, according to the distributed indexing scheme. However, with the recent availability of relatively low cost digital video cameras, a significant growth can be expected for the volume of multimedia content owned by each peer. An index can then become necessary even for the content of a single peer and, to remain autonomous, the peer may prefer to have a local copy of his index. It can be expected that end-user software for content-based indexing of images and video will be released in the near future by several editors, using various image or video descriptors and multidimensional index structures.

After significant efforts dedicated to the digitization of multimedia content (e.g. cultural heritage, scientific documents, etc.), many institutional archives now provide facilities for the online exploration of parts of their collections. Access to the digitized content is sometimes free, but in most cases it is controlled and only synoptic representations (thumbnail images, low resolution video fragments) are freely available. To extend the capabilities of the current keyword-based search tools, several large institutional archives attempt to develop or acquire proprietary tools that can provide content-based indexing for their own collections of multimedia content. To optimize its use of resources, every such individual institution would naturally fine-tune its indexing tool to the specific (possibly *ad hoc*) characteristics of the content it provides and of the users it is familiar with. It can be expected that different archives will adopt different content descriptors and multidimensional index structures, depending on the specific content of the archive and on the selected technology provider. Also, an archive may independently decide to replace its current support technology. An important issue for most institutional archives is keeping full control over the index of the content they own. First, many archives would not accept to make public a major part of their index because they consider this index to be a core asset for their activity. Second, in order to be able to control what part of the content it owns is searchable (or "visible") from the outside, an archive should not make public its full index. Note that this concern for "visibility control" (or, more generally, provider-side privacy) may also become significant for user-generated content.

The above description shows that content is likely to be widely distributed among many autonomous providers that employ heterogeneous index structures (and possibly heterogeneous content descriptions) and are concerned by the privacy of their index. Nevertheless, users cannot accept to search a high number of archives independently, using a proprietary search tool for each. They expect *global*, *uniform*, *efficient* and *effective* search capabilities. In this context, the current peer-to-peer model for content-based retrieval of multimedia content, based on a complete global index and on the open distribution of all content signatures, may need to be revised.

To support both the requirements of the providers and the expectations of the users, in a flexible and scalable way, we suggest a general framework based on the publication by every provider, in a consensual format, of representatives associated to those parts of the local index where the provider intends to make content "findable" by others. The representatives published by all the providers serve to build a global distributed index employed for answering content-based similarity queries.

1P2P4mm, June 6, 2008, Vico Equense, Napoli, Italy