Content-Based Multimedia Indexing and Retrieval

Chabane Djeraba
Nantes University, France

This special issue looks at recent developments in one of the most active fields of research in multimedia information systems: content-based multimedia indexing and retrieval. The technology has wide and exciting potential applications given the number of ongoing projects, prototypes, and attempts at standardization. However, it's still a young field, and few users have yet to integrate it accurately into their everyday activity. The next few years will show whether we can turn this technology into solid products.

Internet search engines such as Alta Vista, Google, and Lycos have become popular, because just by typing a few keywords, users have instant access to a vast amount of documents related to their topic of interest. Because digital documents contain more and more audio, images, and video, it's important that new media is equally accessible through these search engines. This turns out to be a difficult problem, however, because extracting information from those types of data requires more sophisticated techniques than parsing a text document for keywords.

Multimedia indexing and retrieval—aimed at solving these issues—has become an active field of research in the recent years, building on existing research themes such as image analysis and speech recognition and developing new ones such as cut detection, scene segmentation, and text extraction. Most research teams are focusing on some of these issues, based on their own specialization. In a few cases, major efforts have been made to integrate and combine these diverse resources into a single coherent system. (Some of the best examples are the Informedia project at Carnegie Mellon University and the CueVideo project at IBM Research.)

However, because many proposed methods only solve specific problems, it's crucial to clearly understand the potential and limitations of each method. The next step is to define standard evaluation resources with which we can compare these methods. Evaluation is far from trivial because many possible criteria exist and adequate data have to be compiled and made available. Many efforts are ongoing in this direction, such as the Text Retrieval (TREC) conference, the image indexing Benchathlon, and the Advanced Research and Development Activity Video Analysis and Content Extraction (ARDA-VACE) program. This comparative approach has proven challenging in other areas such as speech recognition, information retrieval, and language processing, so there's little doubt that these efforts will greatly contribute to the scientific progress in the field.

I hope this article will provide a brief survey of this innovative field as well as introduce the ideas and concepts explored in this special issue. My aim is to clarify some notions raised by this new technology by reviewing some current capabilities and their potential usefulness to users in various areas. The research results in this special issue hold potential for solving the field's technical problems (see the sidebar "Special Issue Articles"). We can further develop these technologies to produce adequate solutions, provide a wide coverage of the issues and their technical solutions, and build robust systems that we can deploy in real-world visual information systems. The basic ideas behind the special issue's articles are the gap between low-level features and high-level semantics and how to bridge the two levels of indexing. The research presented here should be a springboard for further development in this already explosive field.

Features

Initially, the focus of research solutions in multimedia indexing and retrieval was on content analysis and retrieval techniques linked to a specific medium. Researchers have investigated content-based retrieval from nontext sources such as images, audio, and video. More recently, on the basis of medium archive feedback, researchers have started to combine features from various media. They've also started to study the benefit of

1070-986X/02/$17.00 © 2002 IEEE

Special Issue Articles

The articles in this special issue are based on original submissions to the Second European Workshop on Content-Based Multimedia Indexing, which was held in Brescia, Italy, 19–21 October 2001. The four articles fall into four broad categories that reflect the variety of research directions in the content-based multimedia indexing area: image indexing, video indexing, user access and annotation, and content analysis of video material.

In "Unifying Keywords and Visual Contents in Image Retrieval," Zhou and Huang explore an image retrieval system that considers both high-level features (keywords) and low-level features (colors, textures, and structures). The authors propose a seamless joint querying and relevance-feedback scheme based on keywords and low-level features, incorporating keyword similarities. They also propose a pseudoclassification algorithm that learns the term similarity matrix during user interaction. This learned similarity matrix, specific to the data set and the users, could be applied for keyword semantic grouping, thesaurus construction, and soft query expansion during intelligent image retrieval with user interactions.

In "Semantic Annotation of Sports Videos," Assfalg, Bertini, Colombo, and Del Bimbo discuss semantic sports video annotation. They automatically annotate videos according to elements of visual content at different semantic layers. Their video material can include various sports and can be interwoven with nonsports footage. They decompose each video segment into its visual content elements, including foreground and background, objects, and text captions. They also combine several different low-level features with domain-specific knowledge to capture semantic content at a higher level.

Leonardi and Migliorati's article, "Semantic Indexing of Multimedia Documents," focuses on two different approaches to improve semantic indexing of audio–visual documents. The top-down approach is a semantic indexing algorithm based on finite-state machines and low-level motion indices (lack of motion, pan, zoom, and shot cuts) extracted from the MPEG compressed bitstream. The bottom-up approach performs the indexing with Hidden Markov Models. They analyze several samples from the MPEG-7 content set using the proposed classification schemes.

In Mégret and Jolion's article "Tracking Scale-Space Blobs for Video Description," the authors track blobs derived from the scale space to represent low-level motion information in video materials. They consider this method a first step toward motion characterization and event detection in videos. The blobs contain suitable properties that make them adaptable for tracking (compactness, invariance to simple motion, and spatial distribution). Tracked tokens are gray-level blobs computed in a scale-space framework. On the basis of their experimental observations, they propose a blob-tracking algorithm adapted from an interest point multihypothesis tracker.

knowledge discovery in accurate content descriptions, refining relevance feedback, discriminating multimedia repositories, and more generally, improving indexing. The goal is to handle general queries—for example, "find in video tapes young people going inside banks" and "find in image repositories red flowers in parks during Spring." Answering such queries requires advanced approaches that depend on a central element to describe a medium's content: a feature.

Features are the blood of content-based indexing and retrieval. They are the information we extract from a medium, represent in a suitable way, store in an index, and use during query processing. They characterize the medium signatures. We can classify features into a low and high level according to their complexity and use of semantics.

Low-level features

Low-level features (also known as primitive features) such as object motion (for video), color, texture, shape, spatial location of image elements (for both images and video), special events, and pitch (for audio) permit queries such as "find clips of objects moving from the bottom left to the bottom right of the frame," which might retrieve video pieces of objects (for example, a ball) following that specific trajectory. Other sample queries include

- find images with short, thin, white objects in the bottom right-hand corner;
- find images containing red ellipses arranged in a square;
- find images containing yellow regions in the center; and
- find more images that look like this one.

This level uses features that are objective and directly derivable from the images themselves, but it doesn't refer to any external knowledge base.

Low-level features for indexing are generally extracted automatically and computed efficiently and effectively. The most-favorable application fields are those where we can directly apply low-level features. In these cases, queries are restricted

to application experts—for example, searching collections of fish images. Although the technology of shape retrieval might not be perfect, it's already good enough to identify any fish by its shape. Other favorable areas for retrieval by low-level features are crime prevention (including shoe-print, face, and fingerprint identification), architectural and interior design (retrieval of similar previous designs), medical diagnosis (retrieval of cases with similar features), trademark registration, identifying drawings in design archives, and color matching fashion accessories.

The majority of content-based indexing techniques support low-level features. However, the usefulness of low-level indexing in more general multimedia repositories, such as video databases, art galleries, and museums is still an open problem. For example, in the early years of content-based indexing and retrieval, expectations were high that the technology would efficiently and effectively retrieve images and video pieces from digital libraries, eliminating or at least strongly reducing the need for manual high-level indexing. Disillusionment set in as the realization spread that the techniques under development were of little use for retrieval by semantic content. Video databases now base their retrieval systems on manual high-level features (based on a thesaurus), although a few are experimenting with low-level features in indexing and retrieval software as adjuncts to high-level features. In such applications, there's little firm evidence that current low-level indexing techniques are adequate for multimedia repository exploration tasks. So, it's difficult to present queries that find red flowers in parks during Spring in digital repositories or young people going inside banks in video archives solely based on the low-level features.

High-level features

High-level features (also known as logical, derived, semantic features) involve various degrees of semantics depicted in images, video, and audio. High-level queries are also called semantic queries.

We can distinguish between objective and subjective features. Objective features concern object identification in images and action in video. They permit queries such as "find video clips with a shuttle flight launch during daylight," "find video clips that contain a Ferrari car," "find images that contain Ferraris," and "find a picture of the Arc of Triumph." To answer queries at this level, the retrieval process normally requires prior knowledge. Some prior understanding is necessary to identify an object as a Ferrari rather than a Mercedes and that a given individual structure is the Arc of Triumph. Indexers manually annotate these high-level features. Semantic categories (such as cars, mountains, rivers, plants, buildings, and people) are another facet of the simplest form of high-level features. In this case, users can specify queries such as "find video clips that contain such an action in documentary classes" or "find me more images that look like this and belonging to this class." The semantic classes are generally created manually. These features may consider the outline of the images such as who created the image, where and when, copyrights, authors, and so on.

The subjective features concern abstract attributes. They describe the meaning and purpose of objects or scenes. We can subdivide these features into events (such as independence day), activity types (such as Spanish pop music), emotional meaning (for example, a baby crying), religious (for example, adoration), and so forth. Users can then query these more abstract categories. The retrieval efficiency, on the basis of these features, requires some effort on the part of the searcher and the indexer. Complex interpretation and subjective judgment can be required by an application domain expert to make the relationship between image content and abstract concepts. Features at this level, although less common than low-level features, are in newspaper and art libraries. Operational solutions based on subjective features are rare.

Video features can use temporal relations[1] (such as before, meet, overlap, equal, finish, start, and during) between actions and be objective or subjective. They permit this type of query: "find video clips that contain action 1 overlapping action 2." We can characterize videos not only by fixed images but also by a soundtrack containing music, speech, silence, and other sounds as well as text and graphic objects appearing in a video sequence. All these features allow additional types of queries.

High-level indexing seems a reasonable answer to the semantic drawbacks of low-level indexing. High-level indexing has high expressive power because we can use it to describe almost any aspect of media content. Generally, it's easily extensible for accommodating new notions and can describe media content at varying levels of complexity.

Many available textual retrieval tools can automate the actual search process. However, the process of high-level indexing—whether by key-

words, text, cataloging, or classification—suffers from two important limitations. First, it's inherently time consuming. For example, indexing times can require several minutes per image. Manual indexing times for video are likely to be even longer. Second, manually indexing semantic content isn't particularly suitable for subject retrieval of multimedia document because wide disparities exist in the high-level features that different individuals assign to the same multimedia document. For example, newspapers maintain repositories of fixed images to illustrate articles or documentaries. These repositories contain millions of images and are discouragingly expensive to maintain with detailed, high-level indexing. Broadcasting corporations also deal with millions of hours of video footage repositories, which are hard to manually annotate. It's evident that automatic assistance is necessary.

Low- and high-level feature relationships

Researching both low- and high-level features seems to be a pragmatic way to deal with the shortcomings of current approaches. The techniques of video asset management, which means organization for efficiently reusing video-footage databases, is an obvious example of collaboration between low- and high-level indexing. We can use low-level indexing to break up a video sequence into individual shots and generate representative key frames for each shot. It's therefore possible to generate an entirely automatic storyboard for each video. Even if we use traditional methods to index and classify the video, there can be large time and cost savings.

We can use high-level indexing to annotate key frames. TV companies now use this technology extensively. Current commercial products automatically create storyboards of thumbnail images, which users then manually annotate. We can expect further technology advances, allowing direct search of video content with a much-reduced level of manual annotation, in the near future.

The key to this relationship is automatically constructing high-level semantics on the basis of low-level features. It's one of the biggest challenges of content-based indexing and retrieval and the key to real application achievements. Automatically transcribing text from the speech accompanying video images is one practical way to bridge the gap between low- and high-level features. Other ways researchers have investigated recently are based on three research strategies:

- scene recognition,
- object recognition, and
- knowledge discovery.

The advantages of scene and object recognition include the automatic extraction processes. However, the question remains open about their effectiveness and efficiency with generic applications. Knowledge discovery approaches are semiautomatic with high degrees of automatic assistance.

More precisely, scene recognition consists of identifying the outline scene of images. Some approaches based on colors, textures, regions, and spatial localizations generate texts that any text-retrieval engine can exploit. Others use learning approaches such as neural networks to identify an image's low frequencies or identify color neighborhoods from low-resolution images to generate information according to user-specified knowledge.

Object recognition has been a well-known research field in computer vision for many years. It consists of recognizing and classifying a wide range of objects extracted from a medium (generally fixed images) on the basis of both features of the target objects (region color, shape or texture) and metainformation such as spatial localization, spatial relationships with other objects, and an image's background. Such approaches are based on three simple principles:

- identifying classes of objects,
- depicting image regions that might include examples of the objects, and
- giving an instance of mechanisms to validate the object presence.

Knowledge discovery comes from the data-mining community. It consists of associating extractions between high-level semantics and low-level features from user feedback or medium categorization. One form of knowledge extraction is to select regions from an image, semantically annotate the selected regions, and then apply analog annotations to regions with similar characteristics. We can ameliorate the recall with further user feedback.

Another form of this approach is concept discovery.[3] Different iterations of a query—composed

of motion parameters, color, texture, shape, image example, spatial localization, and significant user feedback—produce a concept composed of query features and instances (relevant clips or images). If the user validates a concept by assigning a semantic label (such as red flowers), the concept and its instances are stored automatically in the database. Over time, a visual thesaurus of concepts is automatically created. Each concept links a description previously labeled by the user, the features associated with it, and its instances. Users may easily and naturally reuse the discovered concepts for future queries, and the concepts are discovered without predetermined information about the application field. However, this approach requires time to build the visual thesaurus and a suitable model to manage the relationships between wide ranges of concepts (for example, overlaps, disjunction, and inclusion).

Another form of knowledge discovery is extracting hidden[2] associations among features (colors and textures) during image indexing. These associations discriminate image repositories. The best associations are automatically selected on the basis of confidence measures. To reduce the combinatory explosion of associations, because repository images contain many colors and textures, these approaches use a visual thesaurus to group similar colors and textures. An algorithm based on a clustering strategy creates the visual thesaurus, which summarizes the image features. The discovered associations contribute to an efficient and effective retrieval process and the automatic classification of images during their insertion into image repositories.

Outstanding issues

Other than the ever-important synergy between low- and high-level features, we must resolve other problems to reveal whether we can turn this technology into solid applications. Some of the outstanding issues include:

- Customizing user queries. Each target application of indexing and retrieval has its own range of special needs and constraints. Systems that fail to address this requirement are unlikely to perform well enough to convince users of the technology's usefulness.

- Identifying best practices in fields that could potentially benefit from content-based multimedia indexing, including management of image collections and drawing archives, electronic publishing, and multimedia content creation.

- Addressing privacy issues and the implications on civil liberties resulting from using images, audio, and video in various applications.

These research and development issues cover a range of fields, many shared with media processing, information retrieval, database technologies, and knowledge discovery.

I hope that the contributions in this special issue provide a stimulant for readers to deal with the problems of content-based multimedia indexing and retrieval. Such contributions are the basis of tomorrow's audio–visual information systems. **MM**

Acknowledgment

Many thanks to Bernard Merialdo (Eurecom, France) and Philippe Joly (IRIT, France) for valuable discussions.

References

1. J.F. Allen, "Maintaining Knowledge About Temporal Intervals," *Comm. ACM*, vol. 26, no. 11, Nov. 1983, pp. 832-843.
2. M. Bouet et al., "Powerful Image Organization in a Visual Retrieval System," *Proc. 6th ACM Int'l Multimedia Conf.* (ACM Multimedia 98), ACM Press, New York, 1998, pp. 315-322.
3. C. Djeraba et al., "Visual and Textual Content-Based Indexing and Retrieval," *Int'l J. Digital Libraries,* vol. 2, no. 4, Apr. 2000, pp. 269-287.

Chabane Djeraba is an associate professor at the Polytechnic School of Nantes University. His research interests include design and analysis of visual information systems, content-based image and audio indexing and retrieval, and multimedia synchronization in an object-oriented database. He has a BS in computer engineering from the National Institute of Computer Science, Algeria; an MSc in computer science from Pierre Mendes France University, France; and a PhD in computer science from Claude Bernard University, France. He is an IEEE member. He is also on the *Journal of Multimedia Tools and Applications* editorial board.