

Multimedia indexing and retrieval: ever great challenges

Chabane Djeraba · Moncef Gabbouj ·
Patrick Bouthemy

Published online: 5 August 2006
© Springer Science + Business Media, LLC 2006

Abstract In this introduction, we present a brief state of the art of multimedia indexing and retrieval as well as highlight some notions explored in the special issue. We hope that the contributions of this special issue will present ingredients for further investigations on this ever challenging domain. The special issue is actually situated between old problems and new challenges, and contribute to understand the next multimedia indexing and retrieval generation. The contributions explore wide range of fields such as signal processing, data mining and information retrieval.

Keywords Multimedia · Information · Indexing · Retrieval · Content · Processing

General Terms Multimedia · Algorithms · Design · Experimentation · Human factors

1 Introduction

Content-based indexing and retrieval is the task of extracting units of information (texts, images, videos, 3D models or a mixture of these) from a collection in order to satisfy some query. The query may take a variety of forms. For example, we may require documents by a particular sketch, colors, motion, person behavior, textures, shapes, author, or about a particular subject. This kind of retrieval has traditionally been achieved by using indexed

C. Djeraba (✉)
LIFL—UMR CNRS-USTL 8022-Building M3, University of Lille 1,
59655 Villeneuve d’Ascq, Cedex, France
e-mail: djeraba@lifl.fr

M. Gabbouj
Tampere University of Technology, Institute of Signal Processing,
P.O. Box 553, FIN-33101 Tampere, Finland
e-mail: Moncef.Gabbouj@tut.fi

P. Bouthemy
IRISA / INRIA Rennes Campus Universitaire de Beaulieu, 35042 Rennes, Cedex, France
e-mail: Patrick.Bouthemy@irisa.fr

metadata that is stored with the units of information. Key features in the metadata may give controlled descriptors to aid the retrieval.

To establish the limits and challenges of content-based multimedia indexing and retrieval, it is useful to start talking about content-based text indexing and retrieval that has years of advances. Content-based retrieval of text is a task that uses the text of information units rather than any added metadata. Free text searching is a good example of content-based text retrieval. The words making up the content of the document are indexed and used as the basis for retrieval. Search engines like Google and AltaVista offer a very good example of content-based text retrieval. In content-based text indexing and retrieval, the process depends on matching content. The textual content of the query is matched with text composing the content of the document, typically organized in suitable structures to accelerate the search process. For text, these search processes are sufficiently well established and widely used to conclude that content-based text retrieval is worthwhile and effective for text information handling. Of course powerful content representation for text retrieval are also widely used. The content matching, on which text content-based processes depend, are in many cases straightforward semantic matches between texts, including powerful numerical vector representations with weights in association with thesaurus, word stemming and textual tricks.

Now let us turn our attention to content based indexing and retrieval with multimedia information (images, audio and video). Can we say with the same high-level user comprehension as we did for text that content-based multimedia indexing and retrieval are worthwhile and effective for multimedia information handling? In short, the answer is “No, certainly not with the same high-level user comprehension.” The main reason is the gap between semantics and multimedia content descriptions. In the following sections we look more closely at content-based multimedia indexing and retrieval techniques, examine some old problems that are ever great challenges. These problems explain why multimedia indexing and retrieval is currently less powerful than text and examine specific efforts to make them more effective.

2 Between old problems and new challenges

Despite interesting advances in content-based multimedia indexing and retrieval, many old problems remain, and are of ever great challenges. One of the key problems is that the signatures (e.g., numerical vectors) of multimedia information are brute and little semantics is being highlighted. Until more powerful multimedia understanding techniques would be developed and integrated into multimedia-content extractions and descriptions, we will be seriously slowed down.

Another problem is the computational problem associated with content-based multimedia retrieval. Many of the representations are multidimensional feature vectors of high-dimensionality and there are serious problems with indexing such features for rapid retrieval. Although novel indexing strategies have been published many of them fail at very high-dimensionality. We believe that significant progress has been realized these recent years, such as in [2, 5].

Finally, it is worth mentioning that multi-modal human—computer interface problems are also associated with content-based multimedia indexing and retrieval. For example, given a query video which contains a complex scene and wishing to use multiple human features (speech words, hand, and visual examples), how do we indicate to the computer the multi-modal queries? How do we synchronize multi-model features for indexing and retrieval.

In the following, we will focus on some ever great challenges: multimedia semantics and user centered systems.

2.1 Multimedia semantics

Content-based multimedia indexing and retrieval has only one objective: finding multimedia semantic information required by the user. Multimedia information does not have a unique semantic, but supports multiple semantics which depend on context, use and experience. Highlighting these semantics and using them in the particular context of content-based indexing and retrieval presents a very exciting challenge for researchers.

Multimedia semantic [1, 4, 6] is both an old problem and a new challenge. It means bridging the semantic gap between primitive features (e.g., object motion, shot intervals, color, texture, shape, pitch, etc.) and semantic primitives (e.g., people, houses, actions, mountains, adoration, etc.). The main disadvantage of information retrieval based on exclusively primitive features (e.g., texture, color, shape) is the little use for retrieval by semantic content. For example, it is not easy to find “President Jacques Chirac in contact with people” in video collections based exclusively on primitive features. Bridging the semantic gap would support various degrees of semantics underlined in multimedia information and permit searches such as “video clips that contain President Jacques Chirac in contact with people.” The heart of the problem is to bridge the semantic gap with reasonable resources, by limiting the manual annotations. For example, annotating Encyclopedia image collections requires 20 min per image.

Automatically extracting text from the speech accompanying video or recognition of text written on images represent one pragmatic way to bridge the semantic gap. Another pragmatic way to bridge the semantic gap is to automatically extract primitive features (key frames, shots, and other classical primitive features) of large video databases, and manually annotate semantic features of a subset of video database. Then, on the basis of the frequent pattern between primitive and semantic features of the subset of video, we generalize the annotations to the remaining multimedia database.

A usual way to cope with the semantic gap is to introduce a learning stage to map numerical features (like image or video features) into semantic concepts (like indoor/outdoor scenes, specific highlights in sports games,...). A major issue in the supervised class learning stage is then the high appearance variability that a given object or a given event may exhibit. Therefore, we have to deal with heterogeneous classes while classes may be not so distant from each other. For a given class, observations and consequently computed image or video features, may vary according to the way the scene is filmed (camera motion, distance to the scene, illumination conditions) and the considered instance of the object or event of interest. This is the case for example in sports video analysis where a class of a given “play” event is actually reflected by several clusters in the feature space of low-level descriptors. Another concern is that a very large amount of labeled data is required to achieve such a goal. It means that a considerable and tedious manual processing has to be performed to build the training video database with ground-truth. As a consequence, active research work is currently conducted to alleviate this problem, for instance by designing semi-supervised recognition methods or by using relevance feedback methods.

Relevance feedbacks is another way to bridge the semantic gap. On the basis of user feedbacks, the system learns the needed user concepts. The disadvantage of such approach is the weakness of the training example set. Learning step is based on few training examples, that are not enough to get a high confidence. What about good examples hired, because not returned by the system? This is the major shortcoming of the classical machine

learning approaches, where the training examples are poor and not representative of the database.

If we consider extensions of machine learning that consider specific data parameters (availability of large reservoirs of data, high dimensions, quality of data, rough, noisy, uncomplete data and features), that are intrinsically dependent on indexing applications, then these extensions are another interesting way to bridge the semantic gap. These extensions are called data mining.

In all cases, ingredients of machine learning principles are interesting to contribute bridging the semantic gap. Machine learning is concerned with the development of techniques which allow computers to “learn”. Machine learning is declined into two forms: unsupervised and supervised.

In the first one (supervised learning), we would like to index semantically videos. So, supervised learning creates a discriminative function from training data (e.g., Primitive features of videos). The training data consist of pairs of input objects (typically vectors), and desired outputs (semantic features such as key-words, associated to videos). The output of the mechanism can predict a class label (semantic class) of the input object (called classification). The task of the supervised learner is to predict the semantic feature of the function for any valid input primitive feature after having seen only a small number of training examples (i.e., pairs of input and target output). To achieve this, the learner has to generalize from the presented data to unseen situations in a “reasonable” way. So, one has to consider various steps:

Determine the type of training examples. Before doing anything else, the expert should decide what kind of data is to be used as an example. For instance, this might be publicity videos pre-classified by an expert in semantic categories: publicity of goods, clothes, geographical regions, cars, people, etc.

Gathering a training set. The training set needs to be characteristic of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.

Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should be large enough to accurately predict the output. As features, we consider both primitive features such as Harris or sift descriptors, wavelets, autocorrelogram, anglogram, etc., and semantic features composed of sample of terms extracted from texts, associated to videos or extracted from audio speech.

Determine the structure of the learned function and corresponding learning algorithm that associate primitive features to semantic features. For example, the expert may choose to use neural networks, decision trees, associations, Bayesian statistics (simplest and efficient approach), SVM, boosting techniques.

Complete the design. The expert then runs the learning algorithm on the gathered training set. Parameters of the learning algorithm may be adjusted by optimizing performance on a subset (called a *validation* set) of the training set, or via cross-validation. After parameter adjustment and learning, the performance of the algorithm may be measured on a test set that is separate from the training set. We input any video publicity and the system extract primitive features and returns the most probable semantic terms associated to video with confidence measure.

In the second one (unsupervised learning), we would like to cluster publicity videos in semantic classes (publicity of clothes, goods, cars, people, services, etc.) with non-a priori

output. Unsupervised learning is a method where a model is fit to observations. It is distinguished from supervised learning by the fact that there is no known output. In unsupervised learning, a data set of input objects (video data) is gathered. Unsupervised learning then typically treats input objects as a set of random video. A joint density model is then built for the data set. Unsupervised learning can be used in conjunction with Bayesian inference to produce conditional probabilities for any of the random videos given the others. Another form of unsupervised learning is clustering, which is sometimes not probabilistic. If the features used to cluster videos are exclusively primitive, then it is not evident to fit semantic clusters. However, if used features are both semantic and primitives then the clusters obtained may fit semantics.

2.2 User centered multimedia indexing and retrieval systems

The users should be in the center of multimedia indexing and retrieval systems. Therefore, it may be interesting to analyze their usage of multimedia information in order to extract their interest, to complete the understanding of their real requirements, and to contribute to the extraction of semantic features, also known by multimedia knowledge. User centered multimedia indexing and retrieval systems include two aspects: the customization of user queries and the identification of the best practices [3]. In the first part, we consider that each target application of multimedia indexing and retrieval has its own range of specific needs. Methods that fail to address these needs are unlikely to perform well enough to convince users of the method usefulness. In the second part, we would like to identify the best practices of the users that could potentially benefit indexing and retrieval. These practices include frequent patterns of multimedia content retrieving and browsing. All these problems become much complex when considering privacy issues and civil freedom from using multimedia information in different context of applications.

User centered multimedia indexing and retrieval systems needs multimodal interface that save different sources of user interactions with multimedia information: what he sees, what he manipulates (play pause, queries, etc.), where, when, and why not emotions of the user when retrieving multimedia information. All these information concern user behaviors on multimedia content. The analysis of user behaviors requires multimodal input interface based on multiple sensors. It takes the form of statistics such as frequency counts of multimedia accesses, the number of video viewing sessions, duration of video shows, number of video viewing sessions for which durations are more than 20 min, average duration of a video viewing session, average number of commands per minute during video viewing sessions, forward transitions, backward transitions, forward jumps and jump ratio. How the user interacts with multimedia information (eye tracking, emotion, browse and navigation operations). All these information show the interest of the users on multimedia content, and would certainly be complementary to annotation and multimedia content analysis processes.

Extraction of user interest from streams of user actions and emotions on multimedia content is an emerging problem. The growing importance of multimedia in every day life (e.g., movie production) increases automatically the importance of multimedia usage analysis. To deal with the increasing volume of available multimedia information, users of these videos need suitable tools that fully utilize the rich source of information hidden in the user behaviors in large multimedia databases for retrieving and navigating through multimedia.

We believe that centering user in multimedia indexing and retrieval is still a young research area and the explosion of multimedia services and usages will certainly promote

the need of automatic analyzing of user interactions on multimedia content to improve the quality of services. To make this approach operational, sensors should be extended to acquire in real time, through log files, the way the user browses and searches. For example, video players need to be extended by this functionality to save user operations (play, pause, forward, etc.). The evolving of non-intrusive sensors will certainly contribute to this area of research.

3 Special issue articles

The articles of the special issue are based on the original submissions to the third International Workshop on Content-Based Multimedia Indexing (CBMI'2003, held in Rennes, September 22–24, 2003, France). The purpose of the workshop was to bring together researchers, developers and practitioners from academia and industry to discuss challenges in indexing and retrieving by content of multimedia information. In this special issue, we sought contributions from a wide range of theoretical and application areas.

When selecting the contributions to be presented in this issue, we aimed at providing a good balance of research areas and contributions to content-based multimedia indexing and retrieval. The 12 papers presented in the issue were selected from a total of 58 papers that have been presented in the Content-Based Multimedia Indexing workshop proceedings, and that had been selected from a total of 90 original submissions to the workshop.

The selected articles fall into broad categories that reflect the variety of research directions in multimedia indexing and retrieval. The purpose of this special issue is to present the current state of the art in the domain covering a broad range of issues.

The special issue is composed of two volumes. The first one, composed of six papers, investigates video and audio indexing and retrieval. The second one, composed of five papers, investigates 2D and 3D images, and interfaces for indexing and retrieval.

The first volume—Video and Audio is composed of:

- Video Retrieval of Near-Duplicates using k-Nearest Neighbor Retrieval of Spatio-Temporal Descriptors, by Daniel F DeMenthon et al.
- Object-based MPEG-2 Video Indexing and Retrieval in a Collaborative Environment, by Michael G. Strintzis et al.
- Information Theoretic Framework for Temporal Segmentation of Videos and Applications, by Bruno Janvier et al.
- Audiovisual Integration for Tennis Broadcast Structuring, by Ewa Kijak et al.
- Audio Indexing: Primary Components Retrieval, by Julien Pinquier et al.
- The Cuidado Music Browser: an End-To-End Electronic Music Distribution System, by François Pachet.

The second volume—2D and 3D Images, Interfaces is composed of:

- An Adaptive Technique for Content-Based Image Retrieval, by Jana Urban et al.
- Content-based Retrieval of 3D Models through Curvature Maps: a CBR Approach Exploiting Media Conversion, by Pietro Pala et al.
- DAG-based Visual Interfaces for Navigation in Indexed Video Content, by Anthony Don et al.
- In Depth Analysis and Evaluation of Saliency-based Color Image Indexing Methods Using Wavelet Salient Features, by Christophe Laurent et al.
- Mental Image Search by Boolean Composition of Region Categories, by Julien Fauqueur et al.

Acknowledgments The evaluation process has involved three review stages to update the papers. We are grateful to the authors for their hard work and patience during this long period of the special issue preparation. The special issue would not be possible without the great effort and reactions of the authors and reviewers.

Here is the list of the program committee members involved in the special issue reviewing process:

Régine André-Obrecht (IRIT, Toulouse), Michel Barlaud (I3S, Sophia-Antipolis), Jenny Benois-Pineau (LABRI, Bordeaux), Catherine Berrut (CLIPS-IMAG, Grenoble), Frédéric Bimbot (IRISA, Rennes), Nozha Boujemaa (INRIA, Rocquencourt), Jean-Cedric Chappelier (EPFL, Lausanne), Chabane Djeraba (LIFL, Lille), Chalopathy Neti (IBM Research), Arjen P. De Vries (CWI, Amsterdam), Alberto Del Bimbo (University of Florence), Nevenka Dimitrova (Philips Research), Gregory Grefenstette (Clairvoyance Corp., Pittsburgh), Patrick Gros (IRISA, Rennes), Benoît Huet (Eurecom, Sophia-Antipolis), Ebroul Izquierdo (QMUL, London), Jean-Michel Jolion (INSA, Lyon), Philippe Joly (IRIT, Toulouse), Riccardo Leonardi (University of Brescia), Benoît Macq (UCL, Louvain-La-Neuve), Stéphane Marchand-Maillet (University of Geneva), Ferran Marques (UPC, Barcelona), Philippe Mulhem (IPAL, National University of Singapore), François Pachet (Sony Research France), Fernando Pereira (IST, Lisbon), Ioannis Pitas (AUT, Thessaloniki), Françoise Prêteux (INT, Evry), Gael Richard (Telecom, Paris), Sin'ichi Satoh (NII, Tokyo), Arnold Smeulders (University of Amsterdam), Michael Strintzis (ITI, Thessaloniki), Murat Tekalp (University of Rochester), Nuno Vasconcelos (HP Cambridge Research Lab), Hong-Jiang Zhang (Microsoft Research, Beijing), Mohamed Daoudi (University of Tours), Christophe Chaillou (LIFL, Lille), Nicu Sebe (University of Amsterdam) and Michael Lew (Leiden University).

References

1. Djeraba C (2002) Content-based multimedia indexing and retrieval. *IEEE Multimedia* 9(2):18–22
2. Kiranyaz S (2005) Advanced techniques for content-based management of multimedia databases. PhD thesis, Tampere, Finland, June
3. Mongy S, Bouali F, Djeraba C (2006) Video usage mining. In: Furht Borko (ed) *Encyclopedia of Multimedia*, Springer, 928–935
4. Rowe L, Jain R (2003) ACM SIGMM retreat report on future directions in multimedia research
5. Urruty T, Belkouche F, Djeraba C (2005) Kpyr: an efficient indexing method. *Proc. IEEE International Conference on Multimedia & Expo, Amsterdam, Netherland, July*
6. Zhao R, Grosky W (2002) Negotiating the semantic gap: from feature maps to semantic landscapes. *Pattern Recog* 35(3):51–58, March



Dr. Chabane Djeraba is a Professor of computer science at University of Sciences and Technologies of Lille, France, since 2003. He is a member of LIFL laboratory which is a joint research unit of both French Scientific Research National Center (CNRS) and University of Science and Technology of Lille, and where he head FOX-MIIRE research team on Mining, indexing of multimedia and complex data. His current main research interests include: optimization of multimedia content descriptions and user behavior analysis on multimedia and complex content. He has been an Associate Professor of computer science in Nantes University, 1994–2003. He obtained a Ph.D. in Computer Science from Claude Bernard University of Lyon, France, 1993; a master degree from Pierre Mendes France University, France, 1990; and an engineer degree from Computer

National Institute (INI), Algiers, Algeria, 1989. He organized major multimedia conferences and being co-chair of workshops such as ACM Multimedia Information Retrieval, and ACM Multimedia Data Mining. He has been the Guest Editor of special issues on top multimedia journals (e.g., *IEEE Multimedia*). During the last 15 years, he published a hundred scientific publications in the areas of multimedia indexing and retrieval and mining, including one leading book in multimedia data mining, and he participated to several program committees of major conferences and journals on multimedia (e.g., *International Journal of Multimedia Tools and Applications*, Springer-Kluwer).



Dr. Moncef Gabbouj is a Professor and Head of the Institute of Signal Processing at Tampere University of Technology, Tampere, Finland. From 1995 to 1998, he was a Professor with the Department of Information Technology of Pori School of Technology and Economics, Pori, and during 1997 and 1998 he was a Senior Research Scientist with the Academy of Finland. From 1994 to 1995, he was an Associate Professor with the Signal Processing Laboratory of Tampere University of Technology, Tampere, Finland. From 1990 to 1993, he was a Senior Research Scientist with the Research Institute for Information Technology, Tampere, Finland. His research interests include multimedia content-based analysis, indexing and retrieval; nonlinear signal and image processing and analysis; and video processing and coding. Dr. Gabbouj is a Distinguished Lecturer for the IEEE Circuits and

Systems Society and Chairman of the IEEE-EURASIP NSIP (Nonlinear Signal and Image Processing) Board. He served as Associate Editor of the IEEE Transactions on Image Processing, and was Guest Editor of the European Journal Applied Signal Processing (Image Analysis for Interactive Multimedia Services, Part I in April and Part II in June 2002) and Signal Processing, special issue on nonlinear digital signal processing (August 1994). He was Chairman or member of several conferences and workshops. WIAMIS 2001 and the TPC Chair of ISCCSP 2004, EUSIPCO 2000, NORSIG 1996 and the DSP track chair of the 1996 IEEE ISCAS. He is also a member of EURASIP AdCom.



Dr. Patrick Bouthemy graduated from Ecole Nationale Supérieure des Télécommunications, Paris, in 1980, and received the Ph.D. degree in Computer Science from the University of Rennes, France, in 1982. From December 1982 until February 1984, he was employed by INRS-Télécommunications, Montréal, P.Q., Canada, in the Department of Visual Communications. Since April 1984 he has been with INRIA, at IRISA in Rennes. He is currently “Directeur de Recherche” Inria and head of Vista group. His current main research interests are: statistical approaches for image sequence processing, motion analysis, motion recognition and classification, motion learning, content-based video indexing. He has been or is currently involved in several European projects or networks. He has conducted several projects in direct connection with French industrial partners. Within the

French academic context, he has been in charge of several working groups in Computer Vision or in Multimedia indexing supported by CNRS and the French Ministry of Research. He was Head of “Comite des projets” of Irisa from 1998 to 2002. He has served as member of the program committees of the major conferences in image processing and computer vision. He was Associate Editor of the IEEE Transactions on Image Processing from 1999 to 2003. He is author of about 40 publications in international journals.