Multimedia Content Analysis, Management and Retrieval: Trends and Challenges

Alan Hanjalic[a], Nicu Sebe[b], and Edward Chang[c]

[a]Delft University of Technology, The Netherlands
[b]University of Amsterdam, The Netherlands
[c]University of California, Santa Barbara, USA

ABSTRACT

Recent advances in computing, communications and storage technology have made multimedia data become prevalent. Multimedia has gained enormous potential in improving the processes in a wide range of fields, such as advertising and marketing, education and training, entertainment, medicine, surveillance, wearable computing, biometrics, and remote sensing. Rich content of multimedia data, built through the synergies of the information contained in different modalities, calls for new and innovative methods for modeling, processing, mining, organizing, and indexing of this data for effective and efficient searching, retrieval, delivery, management and sharing of multimedia content, as required by the applications in the abovementioned fields. The objective of this paper is to present our views on the trends that should be followed when developing such methods, to elaborate on the related research challenges, and to introduce the new conference, *Multimedia Content Analysis, Management and Retrieval*, as a premium venue for presenting and discussing these methods with the scientific community. Starting from 2006, the conference will be held annually as a part of the IS&T/SPIE Electronic Imaging event.

Keywords: Multimedia content analysis, multimedia retrieval, multimedia content management

1. MULTIMEDIA CONTENT ANALYSIS: THE NEXT WAVE

After more than a decade of exploding interest in the multimedia content analysis and retrieval[1,2,3] there has been enough research momentum generated that we can finally reflect on the overall progress. The goal has been to develop automatic analysis techniques for deriving high-level descriptions and annotations, as well as to come up with realistic applications. These applications range from home media library organization that contains volumes of personal video, audio, or images, via multimedia lectures archives and content navigation for broadcast TV and video on demand, to advanced automated surveillance applications using multimedia signals as input. The tools have emerged from traditional image processing and computer vision, audio analysis and processing, and information retrieval. In the following we investigate some important trends in the area and emphasize the related research challenges.

1.1. The Need for Paradigm Shifts in Multimedia Content Analysis

A large majority of multimedia content analysis (MCA) techniques proposed so far have been developed in the way that their applicability is limited to narrow application domains only, and their usability outside these domains is not widely extendible [4]. In many cases, the proposed methods do not work even on a test set covering multiple aspects of one and the same target application domain. The origin of this tendency we see in the fact that the domain knowledge used to bridge the semantic gap between the features and semantic concepts is in most cases far too specific. This leads to the solutions that are inflexible, complex and, therefore, impractical. A statement that could be formulated at this place is that we have to start searching for more generic multimedia content analysis principles instead, which - maybe with some limited fine-tuning related to domain/context-specific conditions - are applicable in a much broader scope than what is currently possible. Important implications are not only the likely increase in robustness of multimedia content analysis algorithms that are based on such principles (due to their inherently larger flexibility and stronger foundations),

Multimedia Content Analysis, Management, and Retrieval 2006,
edited by Edward Y. Chang, Alan Hanjalic, Nicu Sebe, Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 6073, 607301, © 2006 SPIE-IS&T · 0277-786X/06/$15

and the possibility for maintaining their high performance across a wider application scope, but also the possibility to reduce the complexity of electronic devices embedding generic content analysis algorithms (e.g., by developing a common hardware setup and/or system architecture usable across various application domains). More flexibility and less complexity can be obtained by using more generic domain knowledge. As this, however, has as a consequence a decrease in the precision of the content analysis performance, innovative solution concepts need to be found that are able to compensate for this negative effect.

In our view, the major obstacle for finding the required innovative MCA solutions is that the approaches nowadays proposed to solve novel problems in a strongly multidisciplinary research field of multimedia content analysis are still to a large extent biased towards its "parental" and more fundamental research fields such as computer vision, machine learning and signal processing. In other words, the content analysis problems are still approached to a large extent from the point of view of available technology ("what kind of signal processing, machine learning, or computer vision tools do I have available, or know most about?") and not from the point of view of what the optimal solution for the posed multimodal problem would be. A typical example of such practice is an overflow of methods for video content modeling and extraction that are based on various sorts of well-known classifiers, involving classical techniques of supervised learning and an abundance of low-level features extracted from audiovisual signals, but where no convincing motivation is given regarding the suitability of the used techniques and features for optimally solving the posed problems. Clearly, the required innovative multimedia content analysis solutions cannot be found without paradigm shifts in how the problems are approached [4, 5, 6]. Realizing such shifts can be seen as one of the main research challenges in the multimedia content analysis field.

1.2. Content Mining and Knowledge Discovery

Many applications of multimedia content analysis can be addressed successfully without employing complex and biased classifier-training processes but the techniques of unsupervised learning instead. Here we refer to the techniques that discover the semantic content structure (e.g. scenes, stories) and semantic content elements (e.g. events) of a general multimedia document in an unsupervised bottom-up fashion, and with minimum possible domain-specific assumptions. Based on such techniques, largely generic, self-learning content analysis systems can be built that require no off-line training and that are applicable across a wide range of domains and scenarios.

Typical steps involved in the development of such systems would include [4]

- the discovery of the basic content elements of a multimedia document (equivalent to words in a text document),
- reducing these elements to the limited set of *key content elements*, which can be seen as *multimedia keywords*, and which are most powerful in characterizing the content of the multimedia document, and
- using the key elements to discover and index video content semantics.

The outcomes of the unsupervised multimodal content analysis methods can serve directly as retrieval items in a "MultiMedia Google", where they can be matched with user queries that are also represented by multimedia keywords. Such solutions would be quite the opposite of the majority of the solutions being proposed nowadays, which are supervised and tuned to a narrow application scenario/domain, which are difficult to scale and which depend heavily on the available training data set. Although some promising initial results in this direction have already been obtained for the case of composite audio [7], the extension of the concepts of (key) audio elements and semantic audio content clusters to the extremely rich content domain of video is an important challenge that still waits to be pursued by our research community [4].

1.3. Automated Content-based Image Annotation

Automatically providing annotation to images appears to be the most valuable, yet the least utilized application. Almost all applications today work with data elements through verbal metadata. It is the least utilized because automatic annotation is far from being even 50% accurate when the number of semantic classes is large (say exceeding 100). Of course, the large search engines scrape text near a picture and use it for metadata which is not very accurate either.

Before commenting on content-based image annotation, it is useful to point out that context metadata is a source of additionally useful metadata [8]. A time stamp translates quickly into the day of the week, month, and season. A GPS coordinate can yield the country, state, municipality, perhaps even the landmark name. Using camera setting information can yield accurate indoor and outdoor descriptors as well as guess on landscape versus close-up. Researchers have shown that combining the time stamp, location, and camera settings along with inquiries to public databases can yield descriptors such as weather, sunset, dawn, night, day, indoors, and outdoors. Surprisingly, there are few context metadata expansion examples that have left the laboratory.

Context metadata can inform content metadata as in aggregation sites where the inputs are from a variety of sources and where data might be structured, but is incomplete. One shopping site with which we engaged, described their growing "miscellaneous" category – product descriptions that could not be categorized by the automatic verbal classifiers because of ambiguous and incomplete data. One can imagine a memory stick for a computer having many of the same descriptions as the computer or a disk drive storage capacity. However, if the image is matched against existing labeled images, then the metadata or title of the closest match has a high probability of being a good descriptor.

For personal image collections, annotated collections are easier to navigate and find photos. Since the picture subjects are overwhelmingly people, face detection and face recognition may provide the foothold for image annotation to leave the laboratory and entering commercial applications. However, although experimentation has shown a relatively high rate of face recognition (70%-90%), face detection drags down the overall accuracy. In other words, experiments show that when a good facial sample is detected, identification of family members is accurate. The problem is that only about one-third of the people pictures have a machine detectable face.

1.4. User, Context, and Task Modeling

One of the important challenges which still need to be fully addressed is the user interface with the multimodal system. Most tasks performed by the user are not isolated from the interrelated conditions but occur in a certain context. Context awareness can be exploited in retrieving the relevant information, adjusting communication with the user, and adapting the user interface to the current scenario. Many design decisions dictate the underlying techniques used in the interface. For example, adaptability can be addressed using machine learning: rather than using a priori rules to interpret human behavior, we can potentially learn application-, user-, and context-dependent rules by watching the user's behavior in the sensed context. Probabilistic graphical models have an important advantage here: well-known algorithms exist to adapt the models, and it is possible to use prior knowledge when learning new models. For example, a prior model of emotional expression recognition trained based on a certain user can be used as a starting point for learning a model for another user, or for the same user in a different context. Although context sensing and the time needed to learn appropriate rules are significant problems in their own right, many benefits could come from such adaptive systems.

1.5. Multimedia Content Analysis and Retrieval at the Affective Level

So far, the research in the field of MCA has mainly targeted the extraction of the *cognitive* content information that is built of the "facts" about the content carried by audiovisual signals, such as the genre, temporal content structure (shots, scenes, sequences, plot points) and spatiotemporal content elements (objects, persons, events, topics, themes). However, what about the task of finding "exciting" parts of a sport TV broadcast, or "funny", "romantic" or "unpleasant" movie segments? What about locating the segments in a surveillance video that "do not feel right"? Compared to the cognitive content information, which is related to what one sees or hears, here we are interested in how one "feels" about the content one sees or hears. In other words, we would like to obtain the information about the feelings, emotions and moods contained in or evoked by a piece of music or a video clip. We refer to the latter as the *affective* content information or, simply, *affect*.

Although the existing solutions for the cognitive MCA have not yet reached the desired level of maturity, the need has already emerged for starting a parallel research effort in this field that targets the extraction of the affective content [9]. The need for affective MCA theory and algorithms stems from the inability of the cognitive MCA principles to adequately address some of the grand challenges in the field, such as personalized music/video recommendation ("I am now in the mood for this type of video/music!"), highlights extraction ("I want 10 minutes of soccer highlights!"), and automated surveillance.

The deficiencies of the cognitive MCA tools with respect to the abovementioned applications are mainly due to the typically supervised nature of the underlying approaches to bridging the *semantic gap* [2]. These approaches usually involve techniques of pattern classification, like Bayesian networks, Support Vector Machines, neural networks and (hierarchical) Hidden Markov models. An application of these techniques to affective content classification clearly requires the prior specification of the affective content categories (e.g. "happy", "sad", "exciting") that are to be searched for in data, which then needs to be followed by training these categories using a suitable training data set. Realizing this in practice, however, is not as straightforward as it may seem. While finding a representative training data set is already a considerable challenge in the cognitive domain, even for reasonably well-defined problems like face detection, this appears to be far more difficult in the affective domain. The main problem lies in the fact that the variety of the content that can appear in "happy", "sad" or "exciting" video clips is practically unlimited. We can describe an acrobatic action of a soccer player as "exciting", but also the parachute jump or a car-chase scene in a movie.

The content diversity problem discussed above further propagates into the problem of feature-based affect representation. While in the cognitive case the features describe the aspects of a real entity, like for instance, the color *red* being one of the features characterizing a red car, little is known about the relations between the features and something as abstract as affect. Which color combination, sound, texture or temporal effect is to be related to "happiness", "disgust" or "fear"?

2. THE NEW CONFERENCE

The new conference *Multimedia Content Analysis, Management and Retrieval* is a part of the annual IS&T/SPIE Electronic Imaging event, held in the technologically booming heart of the Californian Bay Area. It is a successor of the conference *Storage and Retrieval Methods and Applications for Multimedia*, which has been organized yearly (with small changes in the title) within Electronic Imaging for more than a decade, and has been a premium forum for quality papers addressing the research challenges and applications related to multimedia analysis, management and retrieval. In view of many new trends and brave new research directions that have recently emerged from the multimedia community, we believe a fresh start would be more than welcome at this point. Therefore, we chose for a revised conference concept (and title) starting from 2006.

We did our best to maximize the quality of the paper selection process. From the total of 67 submissions, we accepted 26 papers to be presented orally, and 9 papers to be presented as posters. We thank the members of our Program Committee for their high-quality reviews of the submitted papers, and hope you will find the conference program inspiring and thought-provoking.

For the years to come we hope to set the right tone in this conference by urging the submissions of papers that present novel and fresh ideas, question existing paradigms and unwritten rules, and introduce brave new research directions in the fields of multimedia content analysis, management and retrieval. This trend should be continued in the future editions of this conference, which will make *Multimedia Content Analysis, Management and Retrieval* one of the premium meeting places for the scientists working in multimedia-related fields, and also the place where seminal articles in these fields can be presented, discussed and published.

REFERENCES

1. S. F. Chang, *The Holy Grail of Content-based Media Analysis*, IEEE Multimedia, 9(2):6-10, 2002
2. A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, *Content-based Image Retrieval at the End of the Early Years*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349-1380, 2000
3. N. Dimitrova, H-J. Zhang, B. Shahraray, I. Sezan, T. Huang, A Zakhor, *Applications of Video-content Analysis and Retrieval*, IEEE Multimedia, 9(3):42-55, 2002
4. A. Hanjalic, J. Nesvadba, J. Benois-Pineau, *Moving Away from Narrow-scope Solutions in Multimedia Content Analysis*, 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies, London, November 2005
5. G. Wu, E.Y. Chang, N. Panda: *Formulating Context-dependent Similarity Functions*, ACM International Conference on Multimedia (MM), Singapore, November 2005

6. Y. Wu, E.Y. Chang, B. Tseng: *Multimodal Metadata Fusion Using Causal Strength*, ACM International Conference on Multimedia (MM), Singapore, November 2005
7. R. Cai, L. Lu, A. Hanjalic: *Unsupervised Content Discovery in Composite Audio*, ACM International Conference on Multimedia (MM), Singapore, November 2005
8. E. Chang: *EXTENT Combining Context, Content, and Semantic Ontology for Image Annotation*, ACM SIGMOD CVDB Workshop, June 2005
9. A. Hanjalic, L.-Q. Xu: *Affective Video Content Representation and Modeling*, IEEE Transactions on Multimedia, 7(1): 143-154, 2005