

MOVING AWAY FROM NARROW-SCOPE SOLUTIONS IN MULTIMEDIA CONTENT ANALYSIS

Alan Hanjalic¹, Jan Nesvadba², Jenny Benois-Pineau³

¹ Department of Mediamatics, Delft University of Technology, Delft, The Netherlands

² Storage Systems and Applications Group, Philips Research, Eindhoven, The Netherlands

³ LABRI UMR CNRS 5800, University of Bordeaux 1, France

{a.hanjalic@ewi.tudelft.nl, jan.nesvadba@philips.com, jenny.benois@labri.fr}

Abstract

Recent research results in the field of *Multimedia Content Analysis* (MCA) have been marked by an abundance of theoretical and algorithmic solutions covering narrow application domains only. In this paper we analyze this tendency and its origin in more detail and explain why, in our view, this should not be considered “the way to go” in providing easy access to content in multimedia systems and applications of the future. In particular, we concentrate on the case study of digital video, which we see as a straightforward example of multimedia. Through this case study we will discuss the needs and challenges of improving the generic potential of MCA algorithms.

1 Introduction

A large majority of 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. 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. A statement that could be formulated at this place is that we have to start searching for more generic MCA 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 a likely increase in robustness of MCA algorithms that are based on such principles (due to their inherently larger flexibility and stronger foundations), 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 MCA algorithms, namely by developing a common hardware setup and/or system architecture usable across various application domains [1][2][3]. Clearly, the statement formulated above opens many interesting new research questions, such as, what the possibilities and challenges are for finding generic MCA principles and for developing MCA theory and algorithms based on such principles, or how far we can get until we need domain-specific fine-tuning of MCA algorithms.

Due to its rich content and an immense number of diverse applications where it plays an important role, video appears to be an excellent case study to search for answers to questions stated above. We refer to the term “video” in this paper as, inherently, a multimedia data stream consisting of an image series with accompanying audio and text (e.g. closed captions).

In Section 2 we first briefly outline the general problem of *Video Content Analysis* (VCA) and the typical way of approaching this problem, which generally results in narrow-scope VCA solutions. Then, in Section 3, we elaborate on the deficiencies of narrow-scope VCA solutions, and on the need for more generic foundations of VCA algorithms. Such foundations can be obtained only through paradigm shifts in approaching the VCA problems, which we address in Section 4. An alternative and still insufficiently explored way towards more generic VCA principles involves *content mining and knowledge discovery*, discussed in Section 5. Section 6 summarizes our views developed in this paper, and gives our recommendations for future research in the field of MCA.

2 Video content analysis: Using knowledge to infer knowledge

The amount of video content being produced, watched, edited, stored, broadcasted and exchanged in various application contexts is already phenomenal and quickly growing. Nowadays, terabytes of storage capacity on, e.g. CE devices and networks no longer belong to the realm of fiction and first petabyte solutions are already considered [1][4]. Consequently, users of networks are confronted with a content management problem e.g. to retrieve specifically desired AV content within the home-, mobile-, body-, medical-, car- or community-networks. At the same time, through all the recent advances in digital signal compression, transmission and storage technology, video has gained enormous potential in improving the traditional educational, professional, business, communication and entertainment processes.

In view of the above, there is a growing demand for tools capable of extracting the information from digital video signals, which describes the *semantics* of the content

conveyed by these signals. Such tools should enable us, for instance, to easily access the events, persons, objects and scenes captured by the camera, to quickly retrieve our favorite themes from a large music video archive (e.g. a pop/rock concert database) or to efficiently generate overviews, summaries and abstracts of movies, meeting recordings or video educational material. Figure 1 illustrates the envisioned benefits of VCA on the example of an algorithm set capable of recognizing the “chunks” of digital video data showing an “Alpine landscape”, a “news report on topic T ”, the “highlights of a soccer match” or some “suspicious behavior” detected in a surveillance video.

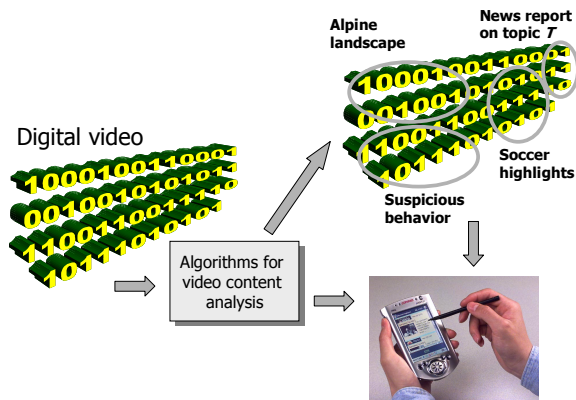


Figure 1. Algorithms for VCA can be developed to reveal the content conveyed by digital video data and to make this content easily accessible to the user.

However, due to high data rates and rich content that is built of the information contained in different modalities (pictures, sound and text) of multimedia documents, the problem of extracting semantics from digital video data streams is considerable. The main difficulty stems from the presence of the so-called *semantic gap*, namely the lack of coincidence between the information (low-level features) that one can extract from the digital data and the interpretation that the same data has for a user in a given situation [5].

A straightforward way of addressing the problem of semantic gap is to use the prior knowledge related to the posed problem in order to establish the missing link between the features and the conveyed semantic meaning. For example, the information that the color of *Ferrari* is red, makes it considerably easier to detect all clips in a Formula 1 video where a *Ferrari* car appears than if this information is not used. Similarly, the detection of goals in a soccer television broadcast may strongly benefit from the availability of the information on the layout of the playfield around the goal port as viewed from a particular (expected) camera position. Furthermore, the Ambient-Intelligence-based sensor networks of future homes not only detect subjects by means of smart cameras [6][7], but also identify and classify these subjects [4]. These networks become more and more reliable over time by means of self-learning and –adapting procedures exploiting updated domain/context information. Not surprisingly, we have

witnessed the omnipresence of what we refer to as *domain knowledge*, in one form or another and at various levels, in virtually all VCA approaches proposed so far [8].

3 From specific to generic domain knowledge

After realizing the great power of domain knowledge in helping the semantics inference from multimedia data streams, a straightforward conclusion when approaching a given VCA problem could be to incorporate as much domain knowledge as possible and so boost the performance. If the incorporated knowledge is too specific, however, this action is very likely to turn counterproductive in the sense that the performance of the resulting VCA algorithms will become highly sensitive to unavoidable small variations in the content being detected. Recalling the example illustrated in Figure 1, the algorithm set embedded into the “black box” will in this case have to be enormous if all possible Alpine landscapes, soccer highlights (and their varying realization and coverage), news reports on topic T covered by numerous different broadcasters, and vast range of “suspicious behaviors”, are to be detected with the same high reliability. We elaborate on this problem in more detail on the example of the well-known problem of detecting highlights in soccer television broadcasts.

Most of the approaches to soccer highlights detection proposed so far are event-based, that is, they aim at detecting pre-defined events that are considered most interesting and therefore usually seen as highlights. Event detection may be approached either by developing feature-based event models (e.g. [9]), by searching for keywords in speech (e.g. [10]), and closed captions (e.g. [11]), by using MPEG-7 metadata (e.g. [12]) or by involving several of the abovementioned clues into inter-modal collaboration (e.g. [13]). While the advantage of this approach is that the detection of each individual event can be optimized using appropriate domain knowledge, its main disadvantage is the need for numerous event models to be able to cover the entire target application scope, namely all soccer television broadcasts. This set of models need to take into account not only all highlight-related events but also various realizations of these events and their coverage that may change from one broadcaster to another. Clearly, the need for numerous reliable (and therefore complex) event models makes the highlights detection process technically and semantically a complex task. Although event detection via keyword spotting may be performed in a more generic way in view of varying event realization and coverage, the resulting set of highlights is limited only to those events, for which obvious keywords are likely to be found. While this may be the case for the “goal” and “penalty” events, other interesting events, such as a nice action in the middle of the field or a nice move of a goalkeeper, will remain undetected.

The problems mentioned above and related to the event-based approach to highlights extraction can jointly be described as the problem of using too specific domain knowledge. An alternative would be to search for a single “generic” event that is assumed to accompany an arbitrary highlighting event, like

for instance, slow-motion segments [14]: interesting events are often replayed in slow motion immediately after they occur. However, moving to a higher level of content representation and understanding – the so-called “affective” level - can lead to even more flexibility and less complexity in detecting highlights. Since, namely, it is realistic to assume that each highlighting event (e.g. goal, penalty, fast action towards the goal, free kick with a goal chance) induces an increase in user’s excitement, we can search for highlights in those video segments that are expected to excite the user most [15]. As argued in [16], the expected variations in a user’s excitement induced by video can be modeled as a function of various content-unrelated features. In this way, the basis is provided for developing generic methods for soccer highlights extraction with no need for pre-specifying the type of highlighting events, and also independent of the event realization and coverage.

4 Needs for paradigm shifts in VCA

The above discussion shows that positioning ourselves a couple of levels higher in terms of analyzing and understanding the domain in which we try to solve a given VCA problem, can greatly improve the usability of the developed solution over the entire target application scope. It would, however, be a bit too simplistic to say that the way out here is in using some more generic domain knowledge only. It is namely so that “more generic” also means “less precise”, the consequence of which is that paradigm shifts may be necessary in the ways we approach the VCA problems. Innovative solution concepts emerging from these shifts, not automatically “biased” any more towards classical computer vision, machine learning and signal processing methods, but developed from scratch and being solidly founded to comply with the real nature of the VCA problem to be solved, have the important task of compensating for reduced precision in the used knowledge to provide similar high performance but with far less complexity, far more flexibility and much wider applicability than the narrow-scope solutions. The need for such flexible solutions can best be seen on the example of complex video genres such as the artistic ones (movies, dramas), where variations in scene realizations can be immense [17]. To better illustrate what we mean by a paradigm shift in VCA, and also to show that such shifts are indeed possible, we briefly elaborate on three of such shifts, which have already been initiated in the field of VCA.

4.1 The principle of content coherence

The first paradigm shift is related to the introduction of the *content coherence* principle [8], a generic principle that enabled the development of widely applicable methods for high-level video parsing. This principle has been introduced in its earliest formulation as *time-constrained clustering* [18] and involves an investigation of the presence of links connecting adjacent video shots and indicating their semantic relation. It has been encouraging to observe the appearance of a vast number of video parsing methods over the years, such

as those based on *content recall* [19], *time-adaptive grouping* [20] or *fast-forward linking* [21], that all aimed and succeeded at further improving the parsing performance while being based on one and the same generic principle. The generic nature of this principle stems from the idea to infer the presence of semantic segment (scene) boundaries at a given time stamp based on the relations between the shots before and after this time stamp. The domain-specific adaptation of this idea can simply be performed by selecting the appropriate features for characterizing and comparing the contents of shots in a particular domain. While visual features appear suitable for parsing movies into episodes, features describing the text statistics and the properties of the speech signal may work best when segmenting news television broadcasts into topics [8].

4.2 From hard-wired to flexible content models

Another important paradigm shift was initiated by the introduction of generic, multi-layer probabilistic models to the field of VCA. The largest impact of this shift is that the introduced models, like for instance the Bayesian Network, removed the need for dedicated, “hard-coded” and therefore inflexible relations between low-level features and higher-level semantic concepts. As can be seen on the example of the semantic inference framework based on “multijets” and “multinets” [22], by employing the abovementioned models it becomes possible to compensate for possible large variations in the content being extracted (e.g. various outdoor, indoor, city or landscape scenes) by spreading the uncertainty over different nodes and layers of the networked model.

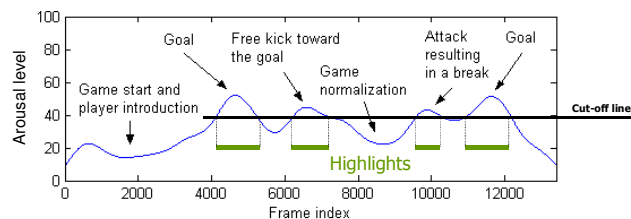


Figure 2. Highlights detection in soccer using affective video content representation. Highlights in the desired length are found by cutting off the arousal time curve at the appropriate level [15]

4.3 VCA at affective level

Finally, we mention here the shift from the cognitive towards the affective level of video content representation and understanding. Recalling the example of soccer highlights detection in Section 3, we can say that through such a shift, a large number of diverse highlighting events spread all over a high-dimensional cognitive content space can be mapped to one “highlights” content class in the two-dimensional affective content space defined by the “arousal” and “valence” axes [16]. This “highlights” content class is represented clearly by high arousal values and therefore easily detectable,

provided that mapping can be obtained between the arousal dimension of the affective content space and low-level features computed from video. Figure 2 illustrates this idea on the example of initial results of highlights detection in soccer obtained using three features and an appropriate model mapping the features onto arousal [15]. Obtaining robust arousal models is, however, still rather difficult and represents an interesting research challenge.

If also a model for another dimension of the affective content space - the valence - is available, we can address something, which seems impossible in the cognitive content space, namely, the detection of “nice”, “funny”, “romantic”, “sad” or “hilarious” moments in video. Or, what about locating “unpleasant” video clips that we would be reluctant to let our children watch? On the first sight, a straightforward approach to extract emotions and moods from video would be to apply the same techniques that are already applied widely in the cognitive domain for object, person, scene and event detection and classification. These techniques are mainly based on various variants of commonly used classifiers 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 of 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. Besides the fact that no training is required, the main advantage of an approach based on the arousal-valence paradigm compared to the approaches based on the classical classification idea, is the possibility to, at least partially, disambiguate the affective states and the related feature sets by reducing the problem of extracting information on affect (emotion, mood) itself to the problem of modeling the less abstract (more comprehensive) affect dimensions. Finding reliable models for valence is, however, a challenge even larger than in the case of arousal [16].

5 VCA with minimum prior knowledge: Content mining and knowledge discovery

The example of affective content modeling has shown that the research efforts in the field of VCA do not necessarily need to follow standard routes of pattern classification. In fact, many applications of VCA 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 MCA 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 include

- 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.

In this way, the outcomes of the unsupervised VCA 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 VCA 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.

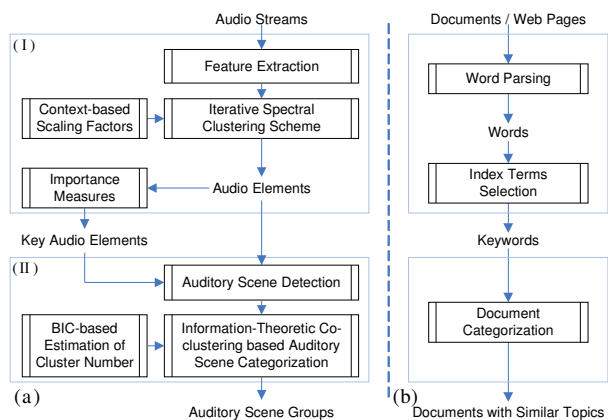


Figure 3. (a) The flowchart of an approach to unsupervised audio content analysis, which consists of two major parts: (I) audio element discovery and key element spotting; and (II) auditory scene categorization. (b) A comparable process of the topic-based text document categorization [23]

As, to the best of our knowledge, no methods exist yet which perform content discovery on a general video data stream, we illustrate the possibilities for developing such a method on the approach proposed in [23] for unsupervised content discovery in composite audio data streams. As opposed to single-modal audio (e.g. pure music or speech), composite audio consists of multiple audio modalities such as speech, music and various audio effects, which are either mixed together or follow each other in a sequence. A flowchart of this approach is given in Figure 3(a). It consists of two major steps: I) audio elements discovery and key audio element spotting, and II) auditory scenes categorization. Both steps are unsupervised and domain- and application-independent. It also facilitates audio content discovery at different semantic levels, such as mid-level audio elements and high-level auditory scenes. This approach can also be seen as an analogy to the topic-based text document categorization [24], as shown in Figure 3 (b). Here, audio elements are the analogies to words, while key audio elements correspond to keywords.

In the method described above, the input is an arbitrary composite audio stream. After feature extraction, an iterative spectral clustering method is used to discover natural semantic sound clusters in the analyzed data stream (e.g. speech, music, noise, applause, speech mixed with music, etc.). These clusters are referred to as *audio elements*. Spectral clustering [25] has proved to be successful in many complicated clustering problems, and proves to be very suitable in this particular case as well. To improve the clustering performance in view of the inhomogeneous distribution densities of various sounds in the feature space, the standard spectral clustering scheme is adjusted by using the context-dependent scaling factors [23]. Using this clustering method, the segments with similar low-level features in the audio stream are grouped into natural semantic clusters that are adopted as audio elements. Then, a number of importance measures are defined and employed to filter the obtained set of audio elements and select the *key audio elements*, which are most prominent in characterizing the content of input audio data.

In the auditory scene categorization step, the potential auditory scenes are first detected by investigating the co-occurrences among various key audio elements in the input audio stream. Then, these auditory scenes are grouped into semantic categories by using the information-theoretic co-clustering algorithm [26], which exploits the relationships among various audio elements and auditory scenes. Moreover, a strategy based on the Bayesian Information Criterion (BIC) is proposed for selecting the optimal cluster numbers for co-clustering.

6 Conclusions and recommendations

In this paper we investigate the tendency in the recent developments in the MCA research area. This tendency is characterized by an abundance of narrow-domain MCA solutions, the usability of which is not only non-extendible to wider application scopes, but in most cases not guaranteed

even on the entire target application scope. 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. 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 VCA performance, innovative VCA 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 VCA solutions is that the approaches nowadays proposed to solve novel problems in a strongly multidisciplinary research field of VCA 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 VCA 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 VCA 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 VCA problems. Clearly, the required innovative VCA solutions cannot be found without paradigm shifts in how the VCA problems are approached. Realizing such shifts can be seen as one of the main research challenges in the MCA field.

Another way of approaching the development of generic MCA solutions is via content mining and knowledge discovery. Instead of making domain-specific assumptions and relying on inherently unreliable training data, the elements of the content structure of a multimedia document could be discovered in an unsupervised, self-learning fashion. Although some promising initial results in this direction have already been obtained for the case of composite audio, the extension of the concepts of (key) audio elements and semantic content clusters to the extremely rich content domain of video is an important challenge that still waits to be pursued by the MCA research community.

7 References

- [1] J. Nesvadba, et al., ‘Real-Time and Distributed AV Content Analysis system for Consumer Electronics Networks’, Proc. Int. Conf. for Multimedia and Expo, Amsterdam, The Netherlands, June 6-8, 2005.

- [2] F de Lange, J. Nesvadba, 'Applying PC network technology to assess new multimedia content analysis applications for future consumer electronics storage devices', 4th Int. Conf. On Intelligent Multimedia Computing and Networking, Salt Lake City, USA, July 21-26, 2005.
- [3] <http://www.multimedien.nl/>
<http://www.extra.research.philips.com/euprojects/candela/>
<http://www.extra.research.philips.com/cassandra>
- [4] J. Nesvadba, P. Miguel Fonseca, R. Kleihorst, H. Broers, J. Fan, 'Face Related Features in Consumer Electronic (CE) device environments', Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics, pp 641-648, ISBN 0-7803-8567-5, The Hague, Netherlands, 2004
- [5] Smeulders A.W.M., Worring M., Santini S., Gupta A., Jain R., 'Content-Based Image Retrieval at the End of the Early Years', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.12, December 2000
- [6] W. Fontijn, J. Nesvadba, A. Sinitsyn, 'Integrating_Media Management towards Ambient Intelligence', Proc of Int. Workshop on Adaptive Multimedia Retrieval, Glasgow, UK, July 28-29, 2005.
- [7] Philips Centre for Industrial Technology, 'Inca 311: Smart Firewire Camera with Rolling Shutter Sensor', <http://www.cft.philips.com/industrialvision>, 2004.
- [8] Hanjalic A., 'Content-Based Analysis of Digital Video', Kluwer Academic Publishers, Norwell MA, ISBN 1-4020-8114-6, 2004
- [9] Gong Y., Sin L.T., Chuan C.H., Zhang H., Sakauchi M., 'Automatic Parsing of TV Soccer Programs', ICMCS '95, pp. 167 -174, May 1995
- [10] Chang Y.-L., Zeng W, Kamel I, Alonso, R., 'Integrated image and speech analysis for content-based video indexing', Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems, 1996, Page(s): 306 -313
- [11] Nitta, N., Babaguchi, N., Kitahashi, T., 'Extracting actors, actions and events from sports video - a fundamental approach to story tracking', Proceedings. 15th International Conference on Pattern Recognition, (ICPR 2000), Volume: 4 , Page(s): 718 -721
- [12] Jaimes A., Echigo T., Teraguchi M., Satoh F., 'Learning personalized video highlights from detailed MPEG-7 metadata', IEEE ICIP 2002, Volume 1, 2002
- [13] Snoek, C.G.M., Worring, M., 'Time interval maximum entropy based event indexing in soccer video', Proceedings of the IEEE International Conference on Multimedia and Expo 2003 (ICME '03), Volume: 3, Page(s): 481 -484
- [14] Pan H., van Beek P., Sezan M.I., 'Detection of slow-motion replay segments in sports video for highlights generation', Proc. IEEE ICASSP, 2001
- [15] Hanjalic A., 'Generic Approach to Highlight Detection in a Sport Video', IEEE International Conference on Image Processing (ICIP 2003), Special Session on Sports Video Analysis (chaired by I. Sezan and B.Li), Barcelona (E), September 2003
- [16] Hanjalic A. Xu, L.-Q.: 'Affective Video Content Representation and Modeling', IEEE Transactions on Multimedia, February, 2005.
- [17] H. Nicolas, A. Manaury, J. Benois-Pineau, W. Dupuy, D. Barba: "Grouping Video Shots into Scenes Based on 1D Mosaic Descriptors", Proc. of IEEE ICIP'2004, Singapore, October 2004
- [18] Yeung M., Yeo B.-L., 'Time-constrained clustering for segmentation of video into story units', Proceedings of the International Conference on Pattern Recognition (ICPR'96), pp. 375-380, 1996
- [19] Kender J.R., Yeo B.-L., 'Video scene segmentation via continuous video coherence', Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 1998
- [20] Rui Y., Huang T.S., Mehrotra S., 'Constructing table-of-content for videos', Multimedia Systems, Special Section on Video Libraries, 7(5), pp. 359-368, 1999
- [21] Hanjalic A., Lagendijk R.L., Biemond J., 'Automated high-level movie segmentation for advanced video-retrieval systems', IEEE Transactions on Circuits and Systems for Video Technology, Vol.9, No.4, pp. 580-588, June 1999
- [22] Naphade M., Huang T.S., 'A probabilistic framework for semantic video indexing, filtering and retrieval', IEEE Transactions on Multimedia, Vol.3, No.1, March 2001
- [23] Cai R., Lu L., Hanjalic A, 'Unsupervised content discovery in composite audio', Proc. ACM Multimedia conference, November 2005
- [24] Baeza-Yates, R., and Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, Boston, MA, 1999.
- [25] Ng, A. Y., Jordan, M. I., and Weiss, Y. 'On spectral clustering: analysis and an algorithm', Advances in Neural Information Processing Systems 14 (Proc. of NIPS 2001), 849-856.
- [26] Dhillon, I. S., Mallela, S., and Modha, D. S., 'Information-theoretic co-clustering', In Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, 89-98.