Towards an integrated personalized interactive video environment

Phivos Mylonas, Kostas Karpouzis, Giorgos Andreou and Stefanos Kollias
Department of Computer Science
School of Electrical and Computer Engineering
National Technical University of Athens, Greece
fmylonas@image.ntua.gr

Abstract

One of the most interesting topics in modern multimedia research, as well as one of the most important tasks in modern audiovisual content providing systems is the treatment of content and users at a semantic level. In this framework, mapping user profiles to multimedia content is a challenging and important problem, as the new generation of home television viewers is currently being confronted with a series of technological developments and improvements, targeted towards their expectations from TV broadcasts. This paper presents initial results from our ongoing work in the field of semantic multimedia analysis, retrieval and user profile extraction in the framework of an interactive video environment, within the MELISA project [12]; this project aims at the cross-media broadcasting of sports events featuring interactive advertising and sports-related games over digital television and provides services for personalized presentation of interactive time video content. Initially, it extends on previous work on low level multimedia content, like scene and shot detection, contour extraction and object tracking, descriptor extraction and matching, and semantic document analysis, in the direction of extraction of semantic user preferences. Such preferences can then be utilized towards the personalization of the overall retrieval process and the multimedia content offering to the end-users. To tackle the latter issue, we propose a methodology based mainly on the utilization of a novel mechanism of weights definition, which are combined with the automatically extracted profiling information.

1. Introduction

In general, a multi-platform content delivery system imposes a number of specific requirements on common media- and communications-related tasks. Specifically in the domain of multimedia content delivery, it is a common fact, that offering and retrieval is by far more difficult to tackle than plain text one, as in this case it is more difficult to match user requests to available multimedia audiovisual content/documents. This is why the role of user profiles in such a task is much more important [6]. The focus of technological attempts in the field of combining user interests with audiovisual archives and multimedia documents can be divided into three major areas: the analysis of a multimedia document for the extraction of the topics related to it, the extraction of user preferences from the analyzed documents and the suitable A/V content offering to the system's end-users.

The problem of analyzing the content of a multimedia document is quite complicated and different than that of analyzing a textual document. Firstly, the entities to be indexed are not directly encountered in the document; recognizable features must be extracted and matched to respective ones found in a knowledge base. Secondly, a multimedia document contains objects and events, whose relations are spatiotemporal, rather than purely grammatical. Furthermore, the extraction and evaluation of users' profiles, based on their usage history habits and preferences stresses the need for intelligent semantic interpretation (e.g. clustering) of the content. Subsequently, suitable user weights are introduced, for the final "balanced" profiling information to be utilized and for usability purposes to be fulfilled.

The structure of this paper is as follows: in section 2 we build upon previous and ongoing research work performed in the fields of object detection [3] and multimedia content analysis [4], utilizing the common meaning and the notion of context in multimedia

documents. Based on these, in section 3 we perform an initial user categorization by extracting user profiles, according to their semantic preferences as well as usage history. The extraction of the user's semantic preferences is performed in a semantic meaningful manner, exploiting the clustering algorithm described analytically in [7], whereas the issue of categorization is further extended by introducing specific weights applied to the system's users. The weights are combined with the previously extracted profiling information to form an integrated A/V entertainment experience for the users. Finally, in section 4 we present the final content adaptation procedure and a notion of our implementation interface of the so far described system, introducing a hands-on scenario, whereas in section 5 we provide our concluding remarks and summarization.

2. Multimedia content analysis

This section of the paper refers to the analysis of the content of a multimedia document with the aim of extracting semantics from it, to be used in the following within the profiling information extraction process. As already mentioned, the main complication when tackling multimedia documents in this direction, shapes in the form of the existence of objects and events, whose relations are spatiotemporal, abstract concepts and events, like "sports" or "goal", which are not explicitly encountered in the content and thus must be inferred, as well as features (like light or distance) which are not attributed to a particular object or event. Several algorithms have been implemented for detecting semantic information using the mpeg-encoded signal of video material. They rely mainly on shot detection and analysis, moving object detection, object feature detection and description extraction and matching with AV description of semantic entities. Scene detection can be considered as the first stage of a non-sequential (non-linear) video representation. For this reason, scene cut detection algorithms [13] are first applied by video indexing and retrieval systems, to extract characteristic frames and shots on which video queries can be applied.

In this work, a hybrid, knowledge – based approach for object recognition in video sequences [3] is used, where objects are modelled, on the one hand, in the signal level through the visual descriptors defined by the MPEG-7 ISO [14] standard for description of audiovisual content; besides this, objects are modelled in the semantic level, through the semantic relations defined by MPEG-7 [8]. This method of video analysis is synopsised initially in the extraction of moving regions using an active contour technique. The proposed method is a contour estimation approach that requires an initial approximation of the position and the size of the objects, successfully dealing with object distortion due to temporal clutter or changes in viewing geometry. Following the task of extracting the bounding polygons and the contours of the detected moving objects in a video segment, MPEG-7 visual descriptors, are utilized to characterize the captured objects or regions. We briefly present and categorize some of them, as follows:

- Color (RGB, HSV, Grayscale)
  - o Variance – An alternative descriptor for the color space, which constitutes a measure for the color variance between the pixels contained in a specified area of a frame.
  - o Percentage/Histogram – The scalable color descriptor.
  - o Spatial Coherency (Structure) – This descriptor captures color content, similarly to color histogram, and its structure.
  - o Spatial Distribution – Specifies the spatial distribution of colors for high-speed retrieval: when applied to a video segment or a moving region, this descriptor specifies the color spatial distribution of a representative frame (or a representative region) selected from the corresponding video segment.
- Shape
  - o Region Shape by ART (Angular Radial Transform) Coefficients – This is a region-based shape descriptor utilizing a 2-D complex transformation defined by a unit disk coordinates.
  - o Curvature – Specifies the contour complexity (smoothness) of an object detected in a video segment.
- Motion
  - o Motion Activity - Captures the intuitive notion of intensity of action in a video segment. The activity descriptor includes the following attributes:
  - o Intensity of Activity – high value of intensity indicates intense activity.
  - o Direction of Activity – the direction element expresses the dominant direction of the activity or categorizes objects according to their motion vectors.
  - o Spatial Distribution of Activity – it is an indication of the number and size of active regions in a frame.
  - o Temporal Distribution of Activity – it expresses the variation of activity over the duration of the video segment.
  - o Spatial Localization of Activity – specifies the number and size of the spatial distribution of motion intensities over the duration of the video by the mean

motion and standard deviation of a moving region in a video segment.

o Rigidity – it is a single-bit flag, which indicates whether an object is rigid or not, according to the direction and the amplitude of its motion vectors.

• Localization

o Region Locator – Localizes a region within images or frames by specifying them with a brief representation of a polygon or box.

o Spatio-temporal Locator – Specifies a frame-by-frame motion trajectory indicated by a chain of region-related positions.

• Edges

o Edge Histogram – Represents the spatial of the edges contained in a frame region; these edges may be directional or non-directional, vertical or horizontal.

Subsequently, matching of these descriptors to ones stored in a semantic knowledge base are used as a means for automatic detection of events and objects. This, of course, remains still a relatively open research issue. Therefore, we introduce a semi-automated manner of tackling this issue, that consists mainly of the existence of an expert, who is needed to recognize the objects that have already been detected and tracked by the system.

As a result, a "semantic index" is generated via recognizing objects and events in a multimedia document and mapping them to "semantic entities" [2]. This issue is also a complicated, still open problem. Similar input, though, can be acquired via textual analysis of the structured textual information contained in the metadata that accompany the annotated multimedia material. So, to summarize, in this work we use a semi-automatic index generation mechanism for multimedia documents, as well as automatic index generation algorithms for textual documents, as the primary step towards the multimedia content analysis.

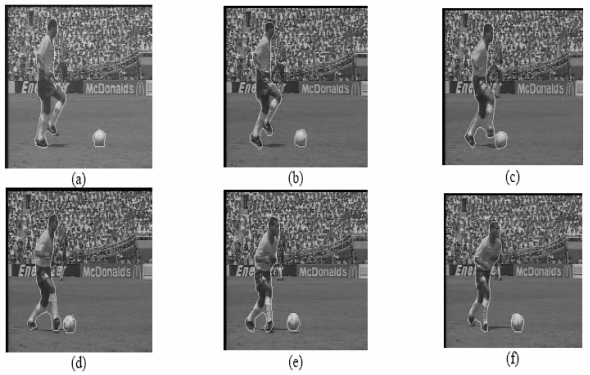

Figure 1. Contour detection in a soccer sequence

Following the initial step, the analysis of the generated semantic index, aims to the extraction of the document's semantics. Towards this goal, a fuzzy knowledge-based approach is followed [4], where each document is represented as a normal fuzzy set on the set of semantic entities. Based on this set, and the knowledge available to the system in the form of semantic relations [10], we detect the degree to which a given document is indeed related to a thematic category. In order to calculate this relation in a meaningful manner, a series of issues are tackled.

1. A semantic entity may be related to multiple topics.
2. A document may be related to multiple topics.
3. The semantic index may contain incorrectly recognized entities.

During the actual process of content analysis we will have to use the common meaning of semantic entities. We refer to this as their "context". A document is represented by its mapping to semantic entities, via the semantic index. Therefore, the context of a document is also defined via the semantic entities that are related to it [10]. So, the "common meaning" of a set of entities, is the context amongst them. In other words it is whatever is common among a set of semantic entities and documents. In this work, we use a combination of 7 fuzzy MPEG-7 semantic relations to provide a fuzzy ordering of semantic entities, i.e. to provide a so-called fuzzy taxonomy, which is ideal for detecting the context: Sp – Specialization, Ct – Context, Ins – Instrument, P – Part, Pat – Patient, Loc – Location, Ag – Agent. This combination of relations forms a (fuzzy, quasi-taxonomic) relation itself and it is used to define, extract and use the context of a document or a set of semantic entities. We define the context of a semantic entity as the set of its ancestors in this combined relation. In Figures 2 and 3, each green bubble represents a semantic entity, whereas each arrow represents a relation between them.

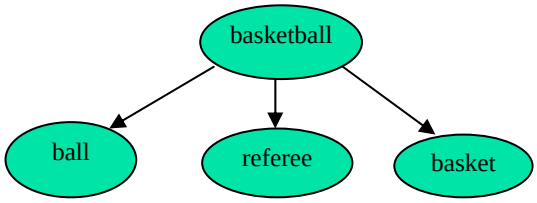

Figure 2. Example semantic entities and relations

Obviously the context is also a set of semantic entities and as more entities are considered it becomes narrower, i.e. it contains less entities and to smaller degrees.

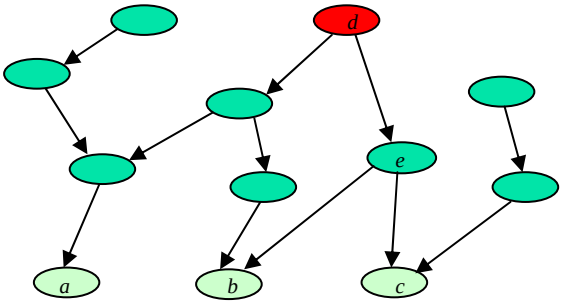

Figure 3. Context of semantic entities

For example, taking into consideration the group of semantic entities that forms the leaves in the tree relation in Figure 3 (i.e. *a*, *b* and *c*), the context of the group, which is again a set of semantic entities, is simply defined by the only common ancestor, *d* on the top. However, assuming only a subset of the leaves (i.e. *b* and *c*) the context is broadened and consists of the set of common descendants of the two semantic entities: the top one, *d* and its right child, *e*, respectively.

Furthermore, a semantic entity can be expanded with its context. For example, the entity "engine" can be expanded to "external combustion". However, if the user asks for documents containing "airplane engine", then obviously an expansion with "external combustion" is not acceptable.

Using the height of the context, which is defined as the greatest value among our semantic entities and demonstrates the degree of relevance, as shown in Figure 4, as a similarity metric (e.g. a measure of the semantic correlation of entities) and applying an agglomerative clustering procedure, we finally obtain the set of topics that correspond to a document; this is the set of topics that correspond to each one the detected clusters of semantic entities that index the given document.

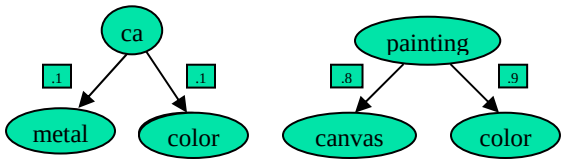

Figure 4. Height of the context

Subsequently, the next step towards a complete multimedia document analysis is formed by the extraction of user preferences, on which the desired extraction of user profiles, accompanied by the user history, will be based. As far as the main guidelines are concerned, the extraction of semantic preferences from a set of documents, given their topics, is quite similar to the above introduced extraction of topics from a document, given its semantic indexing.

3. Content personalization

An important feature that end-users have come to consider as imperative in live sports broadcasts, such as those offered in MELISA, is the provision of visual aids or enhancements. Usually, these enhancements provide quantitative information about specific measurements or events, such as the distance of ball placement to the goal line during a free kick, or a coarse indication of the distance during a long jump in athletics, and can go so far as to display a virtual racing track indicating the real-time positions of cars during a racing event.

Another important aspect that is more closely related to the business model of such an integrated system is that of virtual advertising, where synthetic boards are placed in specific views of the sports venue, displaying content that can be fitted to suit different occasions or even broadcasters. According to the specific content metadata (e.g. the team participating in a football match) or the specific user metadata and profiling (e.g. advertising target groups), different ads related to events or objects are presented to the viewers.

In this context, the need for personalized multimedia services to the end-users is even more stressed. Unquestionably, several types of personalization exist when dealing with interactive services, applications and content delivery to the end-user [1], [5]. This section of the paper refers to "content personalization" based on the underlying automatically generated creation of MELISA's user profiles, where different content can be generated for different individual users or classes of users. Different approaches exist also in the way the system will deal with the gathering of information and statistical analysis. There are two main methods that can be identified: one that relies on the user to provide profile/configuration information, usually via questionnaires, and another where it is the system that gathers the necessary data by filtering usage history information.

In the case of MELISA's users profiles, several issues and different aspects are taken into consideration, while constructing the initial profiles. Main components, on which the profile building process is based, are mainly the specific user needs and interests, the offering possibilities of audio and visual enhancements (e.g. offline vs. real-time in play statistics, online betting functionalities, advertising

functionalities, etc) and several future events alerting functionalities, related to the former information. Moreover, the main hot points to consider, regarding the specific user attitudes, may be summarized in the following:

- A user may be interested in multiple topics.
- Not all topics that are related to a multimedia document in the usage history are necessarily of interest to the user. There is often the case that a document may be accessed occasionally or even more often there is the case a document exists in the usage history, just because the end user accessed it to see that it wasn't of his/her interest at all.

All of the above issues are tackled using similar tools and principles, as the ones used to tackle corresponding problems in content analysis. Thus, once more, the basis on which the extraction of preferences is built is the context. The common topics of multimedia documents are used in order to determine which of them are of interest to the user and which exist in the usage history coincidentally. What is common among two documents, i.e. their common topics, can be referred to as their common context. The height of their common context is used as a metric that can indicate the degree to which two documents are related and subsequently this can be extended to the case of more than two documents, in order to provide a metric that measures the similarity between several multimedia documents. The representation and handling of preferences using fuzzy sets is developed to a greater extent in [11].

Since a user may have multiple interests, we should not expect all documents of the usage history to be related to the same topics. Quite the contrary, similarly to semantic entities that index a document, we should expect most documents to be related to just one of the user's preferences. Therefore, a ranking of documents, based on their common topics, needs to be applied. In this process, documents that are misleading (e.g. documents that the user chose to view once, just to find out that they do not contain anything of interest to him) will probably not be found similar with other documents in the usage history, resulting to only those documents that form the desirable distinct user profile categorization. In order to overcome this obstacle, weights are introduced in order to balance the obtained results from the above user profiling procedure. These weights are initially predefined and stereotyped within the MELISA system specifications from a suitable group of system's usability experts, categorizing its users into three main categories: headstrong users, intermediate users and unconcerned users. As more and more input from every specific end user is gathered, the above procedure results in a weighted mapping of this end user to a specified profile. This mapping changes in that manner continuously and dynamically every user's profiling, until a final equilibrium profile state is achieved. The combination of the profiling information with those weights, results into an ultimate descriptive and representative profile for each user. This final profile is then used in order to provide him/her with A/V content of their interest, as well as with metadata and video enhancements, such as statistical information and explanatory graphs during transmission of the mainstream content (e.g. during sport events).

4. A hands-on scenario

In this section of the paper, we present the final content adaptation procedure, as well as a representative chunk of our system implementation. At this step, all previously extracted information from the usage profiling process is combined with the extra audiovisual enhancements and data accompanied by the system. This content adaptation is performed either because of the large, peculiar nature of the system's end-clients as hardware/software, or because of the specific, special kinds of the enhancements themselves. The latter is clearly based on the content analysis already performed, as well as on the users' usage history manipulation. Links and/or PiP video options to relative "live" or taped sports events form examples of such enhancements. So, to summarize, the final audiovisual content transmission to the end-user evolves continuously according to the supplied extra information.

In MELISA, in order to generate and provide real time content, we use MPEG-4 Rich Media, with audio-visual data of specific sport events, enriched with interactive BIFS graphics. In the initial presentation, the broadcaster/program director embeds all audio-visual objects (images, sounds, etc.) and interactive functionalities (show clickable object, etc.) that will be used in the service. In order to maximize the personalized aspect of the user experience, the initial presentation look-and-feel is optimized for the specific target user terminal. Therefore, separate initial presentations are built for separate user classes, according to their extracted user profiles. Figure 5 shows the initial presentation for an intermediate user profile terminal, regarding a soccer game scenario.

Figure 5. Initial intermediate profile soccer game scene

The interactive functionality that is common to all scenarios targeting end users belonging to this intermediate profile is the Bet Menu depicted on the picture, which allows users to place a bet on a player during the event. When the event starts, the full initial scene together with the first images of the event is sent to the client platforms. During the event, online content is generated to enhance the initial scene with real-time data. The real-time interactive content is encoded as BIFS Updates for the initial MPEG-4 scene. This allows the program director to modify the parameters of the scene at any time, according to what happens in the event. A typical example of real-time content change is the modification of the Bet Menu. Typically the odds for an available bet would vary during the event. Figure 6 shows the result of several generated real-time contents (i.e., BIFS Updates) on the initial presentation given in Figure 5.

Figure 6. Updated intermediate profile soccer game scene

As we see on the top right corner of Figure 6, the score has been changed, as well as the current game time. Also, following an important game action, a new image object has been displayed on top of player 9 which, when selected (by pressing 9 on the remote control) shows statistics for this player. Practically, this corresponds either to the specific user actions per system terminal, or to the specific profile class descriptions. In addition to the BIFS Updates, we also manage a BIFS Carousel, i.e., we periodically generate and send Random Access Points, which consist in merging the initial presentation with the proceeding accumulated updates. The resulting new BIFS scene is useful for users that join the service after the start of the event, so that they provided with up-to-date information, according to their a priori profile categorization and/or terminal/type. Note that although a different initial presentation was used for each target terminal (for example STB or PDA), our framework allows for real-time content scalability. This is achieved by encoding both initial presentations with the same definition for the BIFS Nodes that are supposed to be updated during the event.

Furthermore, from a higher level point of view, several interfaces result from such a (-technically presented above-) combined adaptation, regarding profiling and extra available information. In the following, we additionally present in particular two different screenshots from the simplest available MELISA software terminal type, indicating the differences between the user interfaces, as well as the ones between the content enhancements arriving at the end-user.

Figure 7. Plain client

Figure 8. Enhanced client

In Figure 7 we observe a plain user interface with minimal extra information, whereas in Figure 8 several visual enhancements and interface changes are obvious (like bet, and replay buttons, as well as player number indicator), providing more content management powers to the end–users and indicating the existence of a more enhanced user profile, although the terminal type remains unchanged.

5. Conclusion

This paper is part of our ongoing work in the fields of semantic multimedia analysis and retrieval, personalized multimedia A/V content offering and multimedia system integration. It extended on previous work on low level multimedia document analysis, such as scene and shot detection, contour extraction and object tracking, descriptor extraction and matching and semantic document analysis and semantic document analysis, in the direction of automated extraction of semantic user preferences. Such preferences are utilized together with applicable user weights towards the personalization of the multimedia retrieval process, as well as towards an overall personalized multimedia A/V content experience, concerning the accompanying metadata information and end-user interface. The techniques of this paper are based to a great extend on the utilization of fuzzy relational knowledge representation [9]

Furthermore, the overall MELISA platform [12] introduced in this paper offers a flexible system design for multi-platform interactive content broadcasting. Combination of computer vision techniques, mixed with MPEG-4 based interactive services, such as real-time betting and advertising on clients, provides viewers of sports events a personalized, interactive experience. The methodology, techniques and system presented herein have been developed in the framework of the EU IST MELISA project.

6. Acknowledgments

MELISA is an IST project funded by the European commission in its 5th framework program. The project consortium consists of altogether 11 partners coming from industry and academia. This paper was possible only thanks to the excellent teamwork in the framework of this project. More details can be found under http://melisa.intranet.gr.

7. References

[1] Correia N., Boavida M., Towards an Integrated Personalization Framework: A Taxonomy and Work Proposals, WPTEPW, Malaga, Spain, (2002)

[2] Zhao, R. and W.I. Grosky, Narrowing the Semantic Gap-Improved Text-Based Web Document Retrieval Using Visual Features, IEEE Trans. on Multimedia, Special Issue on Mult. Database, Vol. 4, No 2, June 2002

[3] Tsechpenakis G., Akrivas G., Andreou G., Stamou G. and Kollias S., Knowledge - Assisted Video Analysis and Object Detection, EUNITE, Albufeira, Portugal, September 2002

[4] Wallace M., Mylonas P., Kollias S., Automatic Extraction of Semantic Preferences from Multimedia Documents-WIAMIS, Lisboa, Portugal, April 2004

[5] Spangler W., Gal-Or M., Masy J., Using data mining to profile TV viewers-ACM CACM, December 2003

[6] Angelides, M.C., Special issue on Multimedia content modeling and personalization, IEEE Multimedia 10(4)

[7] Wallace, M., Akrivas, G., Mylonas, P., Avrithis, Y., Kollias, S., Using context and fuzzy relations to interpret multimedia content, CBMI, IRISA, Rennes, France, September 2003

[8] ISO/IEC JTC 1/SC 29 M4242, Text of 15938-5 FDIS Information Technology - Multimedia Content Description Interface - Part 5 Multimedia Description Schemes, 2001

[9] G. Akrivas, G. B. Stamou and S. Kollias, Semantic Association of Multimedia Document Descriptions through Fuzzy Relational Algebra and Fuzzy Reasoning IEEE Transactions on Systems, Man, and Cybernetics, part A, Volume 34 (2), March 2004.

[10] Akrivas, G., Wallace, M., Andreou, G., Stamou, G. and Kollias, S. Context – Sensitive Semantic Query Expansion ICAIS, Divnomorskoe, Russia, 2002

[11] Wallace, M., Akrivas, G., Stamou, G. and Kollias, S. Representation of user preferences and adaptation to context

in multimedia content -- based retrieval, SOFSEM 2002, Milovy, Czech Republic, November 22-29, 2002

[12] E. Papaioannou, K. Karpouzis, P. De Cuetos, H. Guillemot, A. Demiris, N. Ioannidis, MELISA - A Distributed Multimedia System for Multi-Platform Interactive Sports Content Broadcasting, EEE EUROMICRO Conference, August 31st – September 3rd , Rennes, France, 2004

[13] Chong-Wah Ngo, Ting-Chuen Pong, and Roland T. Chin, Video Partitioning by Temporal Slice Coherency, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 11, No. 8, August 2001

[14] ISO/IEC JTC1/SC29/WG11, ”MPEG-7 Overview (v.1.0),” Doc. N3158, Dec. 1999.