Inferring semantics from structural annotations of audio–visual documents

N. Adami, M. Corvaglia, R. Leonardi
DEA - University of Brescia
Via Branze 38, 25123 Brescia, IT
Email: {nicola.adami, marzia.corvaglia, riccardo.leonardi}@ing.unibs.it

Abstract— **In this paper, a new approach for semantic extraction is proposed. Assuming that the semantics of interest associated to a multimedia document is subjective and that the user cannot easily construct a semantic description on different abstraction levels, we propose an interactive tool which allows to generate a semantic description by organizing an audio-visual document. The structural decomposition is the result of a guided annotation by the user: the user segments the input sequence in events, assigns each event to a specific class and includes other informations such as time, place and contained objects. The classification process can evolve dynamically, which means that the user can organize the semantics with various personalized and more specialized classes. Using the resulting structural descriptions and classification, our method automatically generates a richer semantic description. The system is totally MPEG–7 complaint.**

I. Introduction

In modern devices, the large amount of audio–visual material, and the associated metadata, have led many standardization committees to study interoperable means for metadata representation. In September 2001, ISO/IEC standardized MPEG–7. At the same time, many user-oriented tools were created to annotate, classify, summarize, etc. the audio-visual content in a standard compliant framework. One of the first tool was developed by IBM [2], [3], [4]. Similar tools have been made available such as for instance those described in [5], [6] and [7]. All these tools use MPEG–7 for the representation of the multimedia structure in terms of organization of the content into segments. In addition the user can freely annotate each segment but the semantic information, implicitly included, is not used to generate an MPEG–7 semantic description.

One of the reason, might be because the MPEG–7 approach for the representation of semantics is cryptic and not easy to implement automatically. Indeed, some previous works have tried to use the semantic components provided by MPEG–7 in two ways: directly by asking the user to create the semantic description [10] [9] (definition of the semantic entities and of their relationships) or simply asking the user to characterize each entity he/she is annotating [8]. However the first approach is too difficult to use because the user should abstract each concept which is being used while the second approach is too superficial because it allows only a characterization of the semantic entities, without the use of any relationships, abstractions, etc. Moreover additional semantic cues could be inferred from the structural decomposition of the content, aspect not taken into account by the previously mentioned tools.

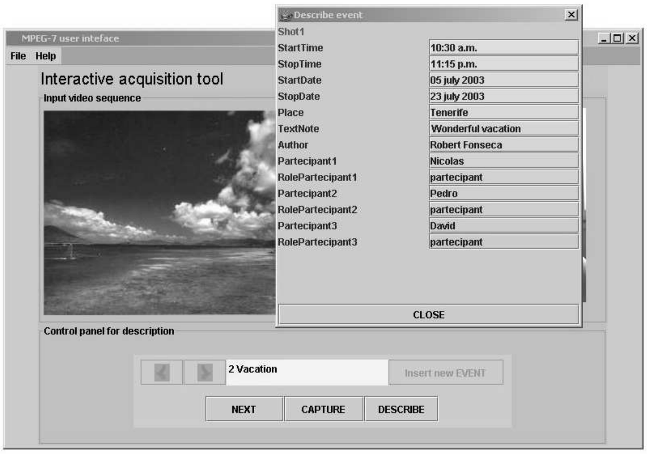
An other reason to fully embrace the development of semantic generation tools is that it is very difficult to formalize and to extract it from the content in a automatic way, mainly because the semantics is subjective. So, the user point of view is crucial. The natural question now is: why don't we create a tool where the user can provide his/her semantics? Or better, we said that the user provides implicitly his/her semantics through an annotation tool as described above, but can further semantics be inferred from the structural organization?

In this work, we try to give an answer to this question. We propose an interactive tool that gives the user many ways to annotate implicitly his/her semantics for a given document (Section III): the user can acquire pictures and videos; the user can then associate to each picture or video segment (shot/scene) information like event type, time, place, people involved; the user can classify the events (*'graduation'*, *'wedding'*, etc.) and the people (*'family'*, *'friend'*, etc.), etc. freely choosing the terms and the hierarchy of the classification . Hence, the resulting MPEG–7 description consists of a segments decomposition where each segment represent a single event (Section III-A and III-B). Then, this structural description can be automatically processed to obtain a richer semantic representation, where semantic entities (events) are linked together using the information provided by the structural decomposition, thus enabling to infer for instance the consecutive happenings of events associated to a certain documents (Section III-C).

II. MPEG–7 descriptions

The aim of the standard MPEG–7 is to provide a set of tools able to describe a wide set of metadata concerning multimedia contents. A MPEG–7 description consists of a set of Descriptors which describe possible representations of the content *features* and Description Schemes which specifiy the structure and relationships between their components, that can be both Descriptors and Description Schemes.

In this work, three types of MPEG–7 descriptions have been used: structural description, semantic description, classification description [1].



Shot1

StartTime	10:30 a.m.
StopTime	11:15 p.m.
StartDate	05 july 2003
StopDate	23 july 2003
Place	Tenerife
TextNote	Wonderful vacation
Author	Robert Fonseca
Partecipant1	Nicolas
RolePartecipant1	partecipant
Partecipant2	Pedro
RolePartecipant2	partecipant
Partecipant3	David
RolePartecipant3	partecipant

2 Vacation

Fig. 1. Interactive tool for descriptions creation.

- The structural description (D_{st}) specifies the structural information given by the multimedia material, which means, for instance, the segmentation in shots with some associated features (visual descriptors, creation and production information, etc.).
- The semantic description (D_{sm}) consists of a set of related semantic entities at a given abstraction level.
- The classification description (D_{cl}) defines one *classification scheme* (CS) which is a set of characteristic key terms given a certain domain pertinent to the document being described (*'graduation'*, *'wedding'*, etc.).

III. Interactive tool

The developed interactive tool (Figure 1) provides to the user an interface to input audio-visual sequences and to annotate them. More in detail, after the acquisition, three main functionalities are available:

- the user can decompose the input sequence in segments, at different levels (segments, sub–segments, etc.) and at leisure;
- the user can associate semantic information, such as time, place, type of event, etc. to each segment at each level.
- the user can record and use new classification terms for characterizing events, places, etc. according to his/her semantic.

The decomposition into segments and their characterization, obtained with the user interaction, lead to the generation of the structural description (D_{st}) and the modification of the classification scheme description (D_{cl}). These two descriptions are processed by the algorithm to generate automatically a semantic description (D_{st}). The system performing the descriptions processing is shown in Figure 2 and it is explained in the following sections.

A. Segment decomposition

The segment decomposition is performed, with the help of the user, in block *User Interface* of Figure 2. This block generates as output the structural description (D_{st}), that is an MPEG–7 compliant segment decomposition. The user can generate the decomposition at leisure, which means that he/she can decompose the input with personal criteria. So, each segment can be seen as an event according with the user point of view. More specifically, the obtained description specifies the structure of the audio-visual data using n different levels of event.

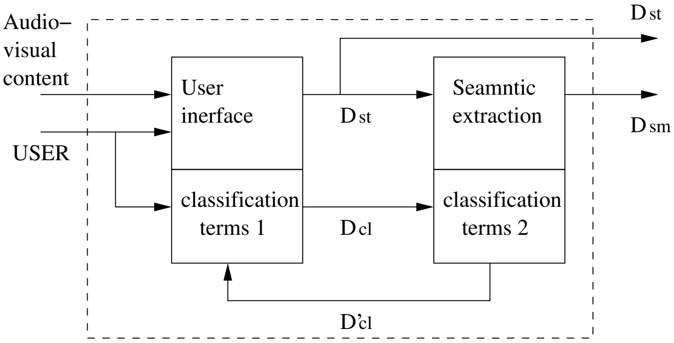


Fig. 2. Interactive tool system for descriptions creation.

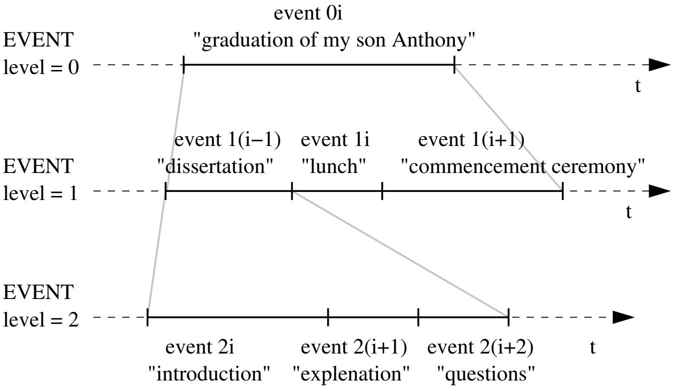


Fig. 3. Structural description.

For instance, we can consider the decomposition of the event ”graduation of my son Anthony” (level zero of decomposition) in two more levels: at level one, we can organize the temporal decomposition in ”dissertation”, ”lunch”, ”commencement ceremony”, etc.; at level two, we can further decompose each segment (event) defined at the previous level. For example, the ”dissertation” can be divided in ”introduction”, ”exposition” and ”questions”. A graphical representation of such decomposition is shown in Figure 3.

This generates a natural hierarchical structure with an implicit semantic associated to each segment, where the relation between segments and events is one-to-one: each event is associated to a unique segment.

B. Segment characterization

The block *User Interface* (Figure 2) provides to the user also tools for the characterization of each event/segment. The event characterization consists of many features, which can be embedded in the structural description (D_{st}), as listed below.

- The user can classify each event, according to individual user preferences. In the considered example, the event at level zero "graduation of my son Anthony" can simply be classified as *'graduation'*, which is a general event;. The events at level one "dissertation", "lunch" and "commencement ceremony" can be respectively classified as *'presentation' 'pause' 'ceremony'*.

- The user can add information to each event (segment):
  - time; for instance, the event "graduation of my son Anthony" can have the additional feature "July" or the event "lunch" the feature "lunch-time";
  - place; for instance, the user can add to the event "lunch" the feature "garden";
  - objects involved (persons, objects, animals, etc.); the user can add other features to the event "lunch", such as "family", "John" and with "Robert".

  For each event, the user can add one or more additional features.
- The user can also classify each additional features: the feature "July" can be classified as *'summer'*, "garden" as *'Outdoor'*, "John" as *'friend'*.

The resulting description leads to hierarchical structure based on the events (Figure 3). Besides, each segment is described by semantic features (event, time, place, objects) according to the user semantic classification. This structural description (D_{st}) has been obtained in a semi–automatic way through a tool that provides the user with the possibility to choose iteratively various semantic components: sequence of the events and the sub–events (segments and sub-segments), semantic features characterizing the events.

As explained above, each event/segment, obtained from the segments decomposition, can be classified referring to a classification description (D_{cl}). The classification description can be predefined or build directly by the user, according to his/her individual preferences. The later alternative represents a good solution for two reasons. First, a predefined D_{cl} cannot represent all possible events and cannot be updated in time. Second, the user should be able to organize such events at leisure considering also a particular context or subjective perspective. In this way, the user can build a custom D_{cl} and directly refer to it.

The classification description consists of a list of key terms at different levels of abstraction. For instance, if we want to classify the periods of the year, at the first level, there are the terms representing the four seasons (*'spring'*, *'summer'*, *'winter'*, *'fall'*); at the second level, there are the terms representing the months of each season (*'summer'* includes *'July'*, *'August'*, *'September'*, etc.); at the third level, the weeks of each month and so on. Hence, we can classify the label "July" with the classification term *'summer'* or *'July'*.

Considering the interface, it has to be noted that the user implicitly creates or updates the classification description (D_{cl}), because in the classification phase he/she can choose a term of classification defined in the past or he/she can introduce a new term in D_{cl} and save the updated D_{cl}, if no suitable terms are available. In this way, the next time the user will use the interface, he/she will have a richer and fully customized D_{cl} . From this point of view, we can say that the user can generate his/her own subjective D_{cl}, that is an ontology associated to the event domain of the document of interest. In Figure 2, we see that available D_{cl} are updated by the user.

C. Semantic extraction

The semantic features considered in previous section are: events, time, place and objects. From a more general point of view, they can be considered *semantic entities*. A semantic description D_{sm} is a set of semantic entities opportunely linked together or, in other words, a semantic description is a graph where each node is a semantic entity and each link a logical relationship between two semantic entities (nodes). So, using all semantic entities, obtained from the user interaction tool, an MPEG–7 semantic description can be extracted.

The semantic entities place, time and objects associated to an event are already linked to the semantic entity event. So, the main task of the *Semantic extraction* block (Figure 2) is to generate semantic relationships among events, which can be extrapolated from the structural description D_{st}: event hierarchy and temporal correspondence between events (*'after'*, *'before'*, etc.), etc. A new graph describing the whole semantic of the original audio-visual material, according to the user's point of view, can be automatically generated (Figure 4).

Consider the structural description generated in Section 3 with the associated features. Analyzing the event "graduation of my son Anthony" at the highest level and the events "dissertation" , "lunch" and "commencement" at the lower level, the following relationships can be derived:

- the semantic entity "graduation of my son Anthony" `contains` the semantic entity "dissertation" , "lunch" and "commencement ceremony" ;
- the semantic entity "dissertation" occurs `before` the semantic entity "pause", which in turn occurs `before` the semantic entity "commencement ceremony" ;
- the semantic entity "dissertation" `contains` the semantic entity "introduction", "explanation" and "questions";
- the semantic entity "introduction" occurs `before` the semantic entity "explanation", which in turn occurs `before` the semantic entity "questions".

The resulting semantic graph is shown in Figure 4. As can be seen, the additional features time, place and objects (semantic entities) are linked to the relative events (semantic entities) by means of well defined relationships:

- the semantic entity "graduation of my son Anthony" is linked to the semantic entity "July" with `time`; with the same link "dissertation" to "10:30 a.m." and "lunch" to "lunch-time";
- the event "lunch" is linked to "garden" by means the relationship `location`;
- "commencement ceremony" is linked to "classmates" with `agent`; likewise, "lunch" is linked to "family", "John" and "Robert".

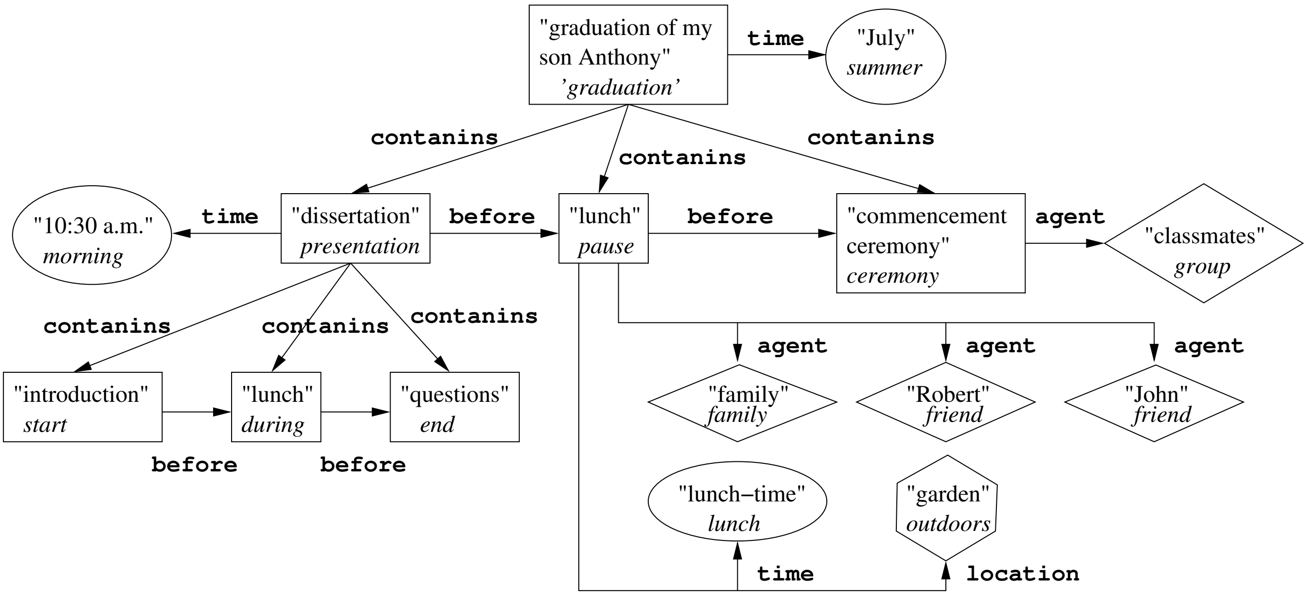


Fig. 4. Semantic description (square = event, circle = time, hexagon = place, rhombus = object).

In the Figure 4, it can noted that each semantic entity is classified; for instance the event ”dissertation” is classified as *’presentation’*. It can be also observed that the semantic description D_{sm} is a graph where semantic entities (events, time, place, objects) are related to each others through well defined MPEG–7 compliant relationships. It is important to stress that the relationships are obtained automatically, according to the structural decomposition and the user classification.

The description of Figure 4 represents a concrete semantics of the actual multimedia content [1]; hence, according to MPEG–7 specifications, the associated abstraction level must be set to zero.

IV. Conclusion and future works

With this work, we propose an application to create three MPEG–7 descriptions: a structural one, a classification one and a semantic one. The first two descriptions are generated with the user's input: the user can decompose input video sequences into events and characterize them with additional information (class information, time, etc.). Moreover, he/she can continuously update an existing classification scheme (CS). Using these two descriptions, the semantic description can be automatically inferred (at a zero level of abstraction): the event classified and the related features are linked together in a graph.

The classification descriptions provide useful information to introduce additional semantic entities in semantic description, like concepts, and to extend semantic descriptions to higher abstraction levels.

In the first case, for example, a new concept can be obtained from the classification of the event at level zero of decomposition (Figure 3): the event ”graduation of my son Anthony” can be classified as ”graduation”, which can further classified as ”life milestones”. This information can be used to generate a new semantic entity of type concept, linked to the semantic entity ”graduation of my son Anthony”.

In the second case, the extension to higher abstraction level can be performed using the classification descriptions information. For instance, using the classification *’graduation’* of the event ”graduation of my son Anthony”, we can easily obtain an abstraction level equal to one, setting the more general event ”graduation of a student”.

Acknowledgment

The authors would like to thank F. Fiorentino for the develop of the interactive user interface prototype.

References

[1] ISO/IEC 15938-5, *Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Schemes*, September 2001.
[2] *http://www.research.ibm.com/VideoAnnEx/index.html.*
[3] Belle L. Tseng, Ching-Yung Lin, and John R. Smith, *Video Personalization and Summarization System*, SPIE Photonics East 2002 - Internet Multimedia Management, Boston, MA, USA, August 2002.
[4] Belle L. Tseng, Ching-Yung Lin, and John R. Smith, *Video Summarization and Personalization for Pervasive Mobile Devices*, SPIE Electronic Imaging 2002 - Storage and Retrieval for Media Databases, San Jose, CA, USA, January 2002.
[5] *http://www.merl.com/projects/video-browsing/.*
[6] *http://www.merl.com/papers/TR2003-34/.*
[7] G. Kazai, M. Lalmas, A. Pearmain and M-L Bourget, *Searching annotated broadcast content on mobile and stationary devices*, IADIS International Conference in Applied Computing, Lisbon, Portugal, March 2004.
[8] Ahmet Ekin, A. Murat Tekalp and Rajiv Mehrotra, *Object-based motion description: from low-level features to semantics*, Proc. SPIE Vol. 4315, p. 362-372, Storage and Retrieval for Media Databases 2001.
[9] Mathias Lux, Jutta Becker and Harald Krottmaier, *Semantic Annotation and Retrieval of Digital Photos*, Forum at the 15th Conference on Advanced Information Systems Engineering (CAiSE 2003), page 85-88, June 2003.
[10] Werner Bailer, Harald Mayer, Helmut Neuschmied, Werner Haas, Mathias Lux and Werner Klieber, *Content-based video retrieval and summarization using MPEG-7*, Proc. Internet Imaging V, pp. 1-12, San Jose, CA, USA, Jannuary 2004.