TRANSFORMING MULTIMEDIA STRUCTURAL INFORMATION INTO SEMANTICS

N. Adami, M. Corvaglia and R. Leonardi

University of Brescia - DEA
Via Branze 38, 25123 Brescia IT
{nicola.adami,marzia.corvaglia,riccardo.leonardi}@ing.unibs.it

ABSTRACT

In this paper a new approach to metadata production is presented. For this purpose, a new interactive tool for audiovisual content acquisition and classification has been developed. The user can decompose a given content into units and easily annotate each unit adding basic information such as time, place, etc. as well classification information such as event type, relationship type, etc. according to the MPEG–7 Standard. At the end of this production process, the tool automatically produces a structural description of the overall set of the annotated units. The new idea proposed in this work is to combine the intrinsic semantics of each annotated unit with the implicit semantic information derived from the structural description, hence reducing the needs to perform complex signal processing operations on the content. This aspect is really important since image and video processing is generally heavier than metadata processing and the content can be spread over a network and not made readily available at the processing point.

1. INTRODUCTION

Audiovisual content analysis aims at the identification of new technologies for enabling a fast and easy access of useful information associated to multimedia documents. Despite of the large number of works carried out in the last decade, nowadays only low level features of the content can be extracted in a fully automated way. On the other hand, only few and with limited capabilities annotation tools have been developed in order to create semantic information, which can be considered as high level features.

Two issues are clear in this scenario. First, no tools are now available for automatic extraction of high level features and even no semantic tools are now able to effectively interact with the user for the generation of semantic description. Second, the gap between the content information we can provide to the user and his/her expectation, the so called "semantic gap" is still present [9].

Hence, our idea is to study a new methodology for automatic extraction of semantic descriptions or at least to develop a new interactive tool where the user can easily interact with the multimedia content through annotation, classification, etc. and then automatically extract meaningful semantic descriptions.

In this work, we developed a new interactive tool following two main guidelines. The first one is to have a more significant user interface providing more functionalities with respect to the most used and popular tools (*VideoAnnEx*, *Caliph&Emir*, *Ricoh Movie Tool*, etc.). The second one, which is the core of this work, is to try to reduce the semantic gap providing to the user some easy mechanisms to express his/her semantics and then formalizing, possibly in an standard compliant, all the created information. In this way the gap will be slightly reduced because the high level features will be better defined and identified for each multimedia content which can also be specified by low level features automatically extracted.

More in detail, the proposed interactive tool (*UniBs.SemTool*) implicitly asks the user to set simple semantics annotation and then automatically generates a richer semantic description. To obtain this semantic description, the user can select a collection of pictures and/or videos and decompose them identifying events (segments) at different hierarchical levels. Semantic properties can further be associated to each segment (events, places, time and objects) while events and objects can be further classified into existing or novel classes. In this way, two first outputs are generated: a structural description and a classification description. These two descriptions can be used to automatically create a third description, which is a semantic description. The latter description is inferred by linking all described events (segments of the structural description) as nodes of a graph using information available in the structural decomposition and in the classification. The role of new classes defined by the user is fundamental. This represent a mean to fix semantic entities and contextual information while providing the user with full control in the generation of the semantic associated with the description. The resulting semantic description is, as it should be, subjective as he/she annotates his/her semantics guided by the interactive tool.

The formalization of the content information provided by the user (structure and classification) and of the inferred

ISBN: 952-15-1364-0

semantic information is MPEG–7 compliant. For the development of the interactive tool MPEG–7 compliant, a Java library has been implemented (*unibs.mpeg7l*).

The rest of the document is organized as follows. The necessary MPEG–7 notion with a brief explanation of the developed MPEG–7 Java library *unibs.mpeg7l* are outlined in Section 2. Section 3 reports an overview and a comparison of the most important user interfaces already developed in other research activities. The results of this analysis brings to the definitions of the requirements of our user interface, whose functionalities and architecture are described in Section 4. Section 5 deals in details with the generation of the content structural description and the automatic extraction of the semantic description. Finally, Section 6 concludes the paper presenting final remarks and future improvement of the proposed methodology.

2. MPEG–7 STANDARD

Given a multimedia content, the broadest Standard used for metadata representation is MPEG–7 Standard. It provides a set of tools and structures able to describe all possible metadata related to multimedia content.

In general, the obvious advantage provided by the use of a Standard is the resulting interoperability between standard elements. In the case of MPEG–7 Standard, we have an additional advantage: MPEG–7 provide a rigorous and interesting formalization to describe both structural and semantic information related to a given content. Therefore, one of the guidelines considered in the definition of the interface (Section 1) can be easily suggested by the Standard MPEG–7.

According to the MPEG–7 philosophy, all metadata can be represented by means of **descriptions**. As many kinds of metadata can be associated to a given content, also many types of MPEG–7 descriptions are available. In our case, we consider three types of descriptions: structural description, semantic description, classification description [10].

- The structural description specifies the structural information given by the multimedia material, which means, for instance, a temporal decomposition in shots with some associated features (visual descriptors, creation and production information, etc.).

- The semantic description consists of a set of related semantic entities limited to a given abstraction level, which means, for instance, the people belonging to a family (entities) with family ties (relationships between entities), such as mother, father, cousin, etc.

- The classification description defines one `ClassificationScheme` (CS) which is a set of characteristic keywords given a certain domain pertinent to the document being described (”mother”, ”father”, ”cousin” or ”vacation”, ”birthday”, etc.).

As it will be explained in detail in next Sections, the user interface will provide to the user easy mechanisms to build a personal temporal segmentation where each segment is characterized and classified by a CS. The obtained MPEG–7 structural and classification descriptions are then processed in order to obtain automatically an MPEG–7 semantic description.

For descriptions processing an MPEG–7 library is required. The only library available is *Mp7Jrs* [2] developed by the Joanneum Research Institute. This library is a C++ implementation for Window, with a consequent reduced range of implementations. Indeed, we developed a Java library *unibs.mpeg7l* which is more portable and better fits the requirements of a user interface.

3. STATE OF THE ART

Recently a few interactive tools have been developed for semantics annotation. We considered only the ones MPEG–7 compliant.

The most famous one is *VideoAnnEx* (Figure 1) developed by IBM [5] [4]. With this tool, the user can choose an input video sequence, decompose it at leisure and annotate each segment [13] [14] [12]. The annotation involves the selection of the classes the segments belong to and writing free text annotation. The classes allow the user to identify each segment with a particular semantics (e.g. ”Outdoors”, ”Indoors”, etc.); the user can use predefined classes or create his/her own classes. Such classes are characterized by a hierarchical structure, so that the classification can occur at different levels, for instance a segment representing the sky with clouds belongs to the class ”Outdoors”, in turn to the class ”Nature (Low Level)” and finally to the class ”Cloud”. When the user annotates each segment, all components of the description are MPEG–7 compliant where each segment is characterized by its temporal references (start point and duration), by a list of MPEG–7 `FreeTextAnnotations` [10] (this number can be freely adjusted by the user) simply reporting the terms of classification.

```
... <VideoSegment>
  <TextAnnotation ...>
    <FreeTextAnnotation>
        Outdoors
    </FreeTextAnnotation>
    <FreeTextAnnotation>
        Nature_(Low-level)
    </FreeTextAnnotation>
    <FreeTextAnnotation>
        Cloud
    </FreeTextAnnotation>
    <FreeTextAnnotation>
        My free annotation
    </FreeTextAnnotation>
  </TextAnnotation> ...
```

This implies that the classification information is stored, but they not exploit the full MPEG–7 potentialities: there are no links between the items of classification and the annotated components of the description. We can say that the classification information provided by IBM tool satisfies only interface requirements and does not attempt in anyway to capture the coherence of the representation created by the user. *VideoAnnEx* provides an other interesting functionality: for each shot, the user can associate to each annotation (classification or free text) a region of the keyframe, by selecting a square of such keyframe.

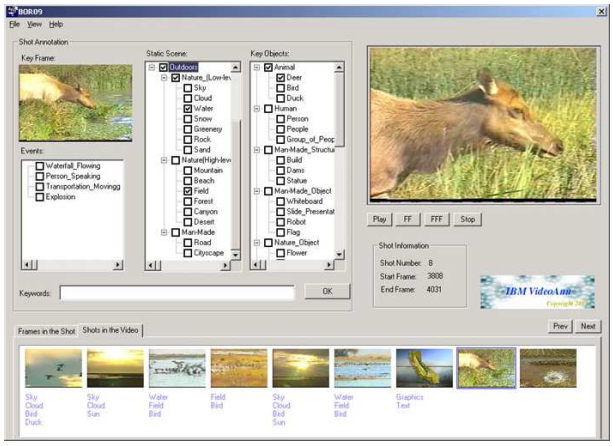

Fig. 1. *VideoAnnEx* interface.

In the context of this paper, the *VideoAnnEx* features of interest are temporal decompositions, even if only at one level (shots decomposition), and the semantic characterization of each obtained segment performed by the user. In the same terms, other tools have been developed. *Ricoh Movie Tool* [6] is an interactive tool where the user can decompose a video at different levels, annotate or classify each obtained segment. In a similar way, *MPEG-7 Annotation Tool* [3], developed by EPFL institute, provides hierarchical temporal segmentation where each segment can be annotated by the user. In all these tools, during the entire annotation process the semantic entities and their relationships, implicitly provided by user, are not captured. Indeed, the user provided unawarely his/her semantics through the decompositions in segments, the free text annotation and the classification. Besides, this last information is not used in order to extract a richer description.

An interesting effort to extract semantics from a given multimedia content is provided by the interactive tool *Caliph&Emir* [1]. As shown in Figure 2, given an image, this tool provide to the user a lot of metadata to be set: free and text annotation, creation information, low-level visual features (`ScalableColor`, `EdgeHistogram`), semantic information. The innovative aspect brought by this tool is the MPEG–7 semantic description generated for the described image. More in detail, the semantic descriptions are built interactively with the user by setting some of the MPEG–7 components (such as semantic entities, relationships, concepts, etc.) [11] [7]. This approach is often difficult for the user as he/she usually has no clear concept in mind or he/she is not concerned by a particular abstraction level. Besides, this tool assumes that the user knows MPEG–7 Standard, but it is not generally true. The only contents that can be described with *Caliph&Emir* are images. The extension for video segments is reported in [8].

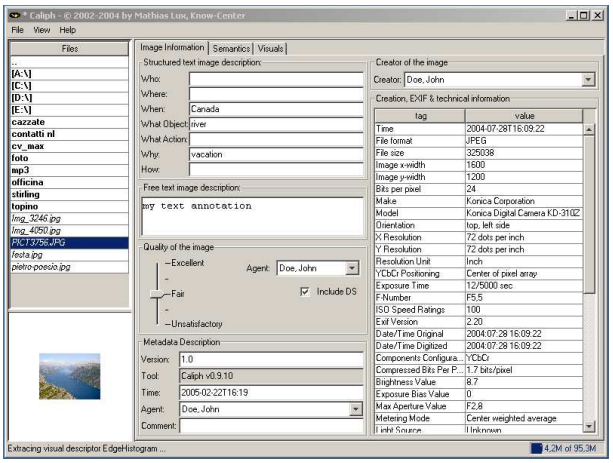

Fig. 2. *Caliph* interface.

Here after, we consider only the two interactive tools *Caliph&Emir* and *VideoAnnEx* because the other tools analyzed have less or similar functionalities. A compared summary of the functionalities of *Caliph&Emir* and *VideoAnnEx* is reported in Table 1.

4. *UNIBS* INTERACTION TOOL

In Section 3, *Caliph&Emir* and *VideoAnnEx* are analyzed and compared (Table 1). From the comparison, we can conclude that they somehow provide complementary functionalities: *Caliph&Emir* describes images while *VideoAnnEx* videos, *Caliph&Emir* produces semantic descriptions while *VideoAnnEx* structural descriptions, etc. From an other point of view, both tools demand a big effort on the user: in *Caliph&Emir*, for instance, the user should be able to build an MPEG–7 semantic description without any help.

Starting from all these considerations and keeping in mind that our purpose is the automatic extraction of semantic descriptions, we defined the architecture of the system and the user interface, following the requirements listed above:

- automatic semantic generation;

Caliph	*VideoAnnEx*
Images description (one content type)	Videos description (one content type)
No structural description (no segmentations)	Structural description (temporal segmentation)
Semantic description	No semantic description
Free or structured text annotation	Free text annotation
No classification descriptions	Classification descriptions (not used!)
Low-level visual features – `EdgeHistogram` – `ScalableColor`	No Low-level visual features
No regions description	Region annotation
Creation information	No creation information
Many functionalities	Few functionalities
Very difficult to use (without knowing MPEG–7)	Easy to use

Tab. 1. Comparison between *Caliph* and *VideoAnnEx*.

- hierarchical temporal decomposition of video and audio contents at many different levels with the generation of a structural description;
- hierarchical organization of the images;
- semantic characterization of each basic element (video segments, images, etc.) by means of semantic annotation and classification provided by the user;
- generation and progressive update of the classification descriptions provided by the user;
- user friendly and simple user interface;
- not explicit reference to MPEG–7 tools and structure;
- description of all possible contents: images, videos, sounds, etc.;

The system implemented is shown in Figure 3. The user can easily select, organize, annotate and classify the content at pleasure (*User Interface*). During this phase, two MPEG–7 descriptions are/or generated: *Structural description* and *Classification description*. Subsequently, a *Semantic description* is automatically produced.

The methodology for the generation of the descriptions is explained in depth in Section 5 while the user interface *UniBs.SemTool* is shown in Figure 4. The user interface consists of four main parts:

- content browser (top–left),
- content player (button–left),

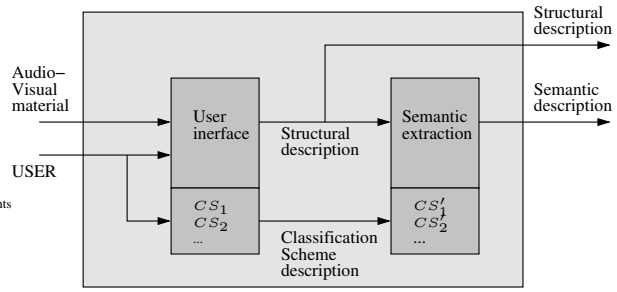

Fig. 3. *UniBs.Semtool* system.

- panel for annotation and classification (right),
- pop-up for video/audio temporal decompositions or image organization (center).

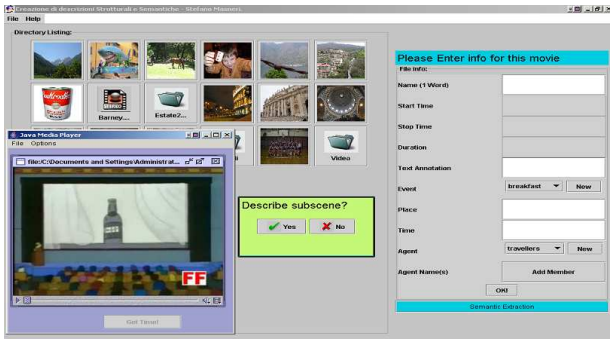

Fig. 4. *UniBs.SemTool* interface.

5. GENERATION OF MPEG–7 DESCRIPTIONS WITH *UNIBS.SEMTOOL*

Referring to Figure 3, we can divide the description generation into two phases: interactive generation of the structural description (Section 5.1) and then automatic generation of the semantic description (Section 5.2). Classification descriptions are first generated and then updated all the times the user requires to add a new term (class) in the CS.

5.1. Decomposition into events

As shown in Figure 3, the structural description is generated by the interactive tool as output of the first block. This description is based on the decomposition of the audio-visual material into events (pop–up in the center of Figure 4) and on the annotation associated to each event by the user (right part of Figure 4). More specifically:

- The description specifies the structure of the audio-visual data using n different "semantic" levels. For instance, if we consider two levels and we want to

describe the summer, at the first level, we can organize the temporal segmentation into weeks of work, vacation, etc. At the second level we can further decompose each segment (event) defined at the previous level; for instance, the vacation can be divided into excursion, dinner, etc. This generates a natural hierarchical structure with an implicit semantic associated to each segment. The relation between segments and events is one-to-one: each event is associated to a unique segment.

- The crucial point for this structural representation is the classification of each event, according to individual user preferences. For this purpose, we can use a predefined CS available in MPEG–7 or we can build a new CS which suits the requirements of our tool. This alternative represents a good solution for two reasons. First, the predefined CSs in MPEG–7 do not represent all the possible events. Second, the user should be able to organize such events at leisure, considering a particular context or subjective perspective as well. In this way, the user can build implicitly the CS and directly use it. Considering the interface, this means that the user can classify an event selecting an item already available in the CS database; if a suitable item is not available in the CS database, the user can define a new item and enrich the CS database. In Figure 3, we see that available CS are updated by the user.

```
<ClassificationScheme
        uri="urn:mpeg:mpeg7:cs:event_lev0">
  <Term termID="birthday">
    <Name xml:lang="en">Birthday</Name>
  </Term>
  <Term termID="vacation">
    <Name xml:lang="en">Vacation</Name>
  </Term>
... </ClassificationScheme>
```

An event can be associated as a semantic entity to a segment, using `SemanticBase` of type `EventType`.

```
<VideoSegment id="Vacation2004-1">
  ...
  <SemanticBase xsi:type="EventType"
                id="E1_vac2002">
    <Label href="urn:mpeg:mpeg7:cs:
                 event_lev0:vacation">
      <Name>Vacation</Name>
    </Label>
  </SemanticBase>
  ...
</VideoSegment>
```

- The user can add information to each event (segment): time, place and objects according to the definition of the semantic tools in MPEG–7. Each of these terms can be seen as a semantic entity (`SemanticBase`), respectively of type `SemanticTimeType`, `SemanticPlaceType`, `AgentObjectType`. We need an additional feature: associate such information to the event. For this purpose, the semantic entity `SemanticBase` of type `EventType` includes the element `Relation`. Using a restricted set of relation types, in this case belonging to the MPEG–7 predefined CS `SemanticRelationCS`, we can associate additional information to each event. In our implementation, a simple set of relations was considered: `time`, `place` and `agent`.

```
<SemanticBase xsi:type="EventType"
              id="E1_vac2002">
  <Label href="urn:mpeg:mpeg7:cs:
               event_lev0:vacation">
    <Name>Vacation</Name>
  </Label>
  <Relation target="#ST1_vac2002"
            type="urn:mpeg:mpeg7:cs:
                  SemanticRelationCS:2001:time" />
  <Relation target="#SP1_vac2002"
            type="urn:mpeg:mpeg7:cs:
                  SemanticRelationCS:2001:location" />
  <Relation target="#AO1_vac2002"
            type="urn:mpeg:mpeg7:cs:
                  SemanticRelationCS:2001:agent" />
</SemanticBase>

<SemanticBase xsi:type="SemanticPlaceType"
              id="SP1_vac2002">
  <Label>
    <Name>Tenerife</Name>
  </Label>
</SemanticBase>

<SemanticBase xsi:type="SemanticTimeType"
              id="ST1_vac2002"> ...
</SemanticBase>

<SemanticBase xsi:type="AgentObjectType"
              id="AO1_vac2002">
  <Label href="urn:mpeg:mpeg7:cs:role:travellers">
    <Name>Travellers</Name>
  </Label>
  <Agent xsi:type="PersonGroupType">
     <Member>
       <Name>
         <GivenName>Robert</GivenName>
       </Name>
     </Member>
     <Member> ...
</SemanticBase>
```

- The user is usually mainly interested to the relation between agent objects, that represent people, animals, etc. involved in what he/she is describing. Using the DS `SemanticBase`, we can follow two directions. In the first case, `SemanticBase` of type `AgentObjectType` provides us three possible types of `Agent`: `PersonType`, `PersonGroupType` or `OrganizationType`. Whatever we choose, we can classify only the whole entity (the person, the group or the association) while we cannot, for instance, describe the relationships between the members of the same group (`PersonGroupType`). To get over this limitation (second case), we can specify each elementary agent object as a semantic entity (`SemanticBase`) and specify the relationships be-

tween them using a CS constructed directly by the user, as it was described for the event classification.

The resulting description leads to hierarchical event based structure (Figure 5). Besides, each event is described by semantic entities (event, time, place, object agents) through simple links. This description is 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 entities characterizing the events, classes (CS) to define events and relationships between agent objects.

To generate an MPEG-7 compliant implementation, the description starts with all the CSs (`Description` of type `ClassificationSchemeDescriptionType`) and then with the description of the decomposition into events using `Description` of type `ContentEntityType`. The list of the `SemanticBase` of each event is included using the element `Semantic`, child element of `Video–Segment` which is used to describe each event or sub–event.

```
<Mpeg7 ...>
  <Description
        xsi:type="ClassificationSchemeDescriptionType">
    <ClassificationScheme
        uri="urn:mpeg:mpeg7:cs:event_lev0">
      <Term termID="birthday">
      ...
    </ClassificationScheme>
  </Description>
</Mpeg7>
```

```
<Mpeg7 ...>
  <Description xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="VideoType">
      <Video>
        <TemporalDecomposition gap="true"
                               overlap="false">
          <VideoSegment id="Vacation2004-1">
            <Semantic>
              <Label/>
              <SemanticBase xsi:type="EventType"
                            id="E1_vac2002">
                ...
            </Semantic>
          </VideoSegment>
        </TemporalDecomposition>
      </Video>
    </MultimediaContent>
  </Description>
</Mpeg7>
```

5.2. Automatic semantic extraction

In the previous section, we analyzed how the structural description and the classification descriptions are generated by the user using the interactive tool. These descriptions are processed in a second time in order to extract automatically a richer semantic description (see second block in Figure 3). For this purpose, we remember that the structural description describes a segmentation of content where each

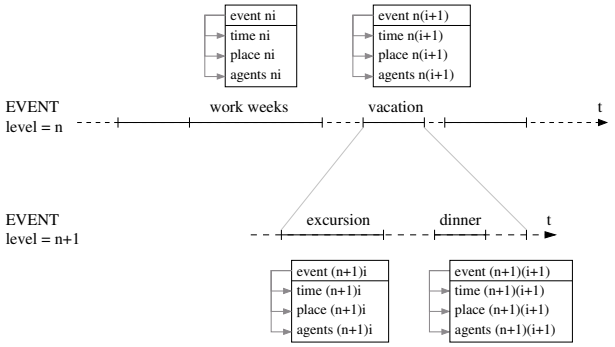

Fig. 5. Structural decomposition into events.

segment corresponds to a semantic entity (event) linked to or containing other semantic entities (time, place, agent objects).

The objective is to build automatically a pure semantic description, that is a graph where the nodes are the events and the links the relationships among them. Starting from the structural description, the semantic entities (events) are already available while the relationships among entities can be inferred by the structure imposed by the interactive tool.

A semantic description in MPEG–7 starts with `Description` of type `SemanticDescriptionType`. The following mandatory element is `Semantics` which is of type `SemanticBagType`. This means that `Semantics` can include a set (”bag”) of semantic entities (`SemanticBase`). In turn, we can use this property to collect all the semantic entities available from the structural description (events, places, times, agent objects) into a simple list. Place, time and agent object entities are already related to the corresponding event, but all the events are at the same level. We can create semantic relationships between events looking at the structural decomposition into events (segments) by using a semantic graph: each node of the graph represents an event and each link the relationships between events.

Consider the example described in section 5.1, that is the structural description about the summer. If we consider the event ”vacation” (node A) at the highest level and the events ”excursion on the first day” (node B) and ”dinner on the forth day” (node B) at the lower level, the following relationships can be derived:

- the event ”vacation” (A) *contains* the event ”excursion” (B) and the event ”dinner” (C);
- the event ”excursion” (B) occurs *before* the event ”dinner” (C).

The resulting semantic graph is shown in (Figure 6).

According to MPEG–7 Standard, we can use the DS `Graph` (one of the possible child element of `Semantics`): `Node` will be set with the events ”vacation”, ”excursion”,

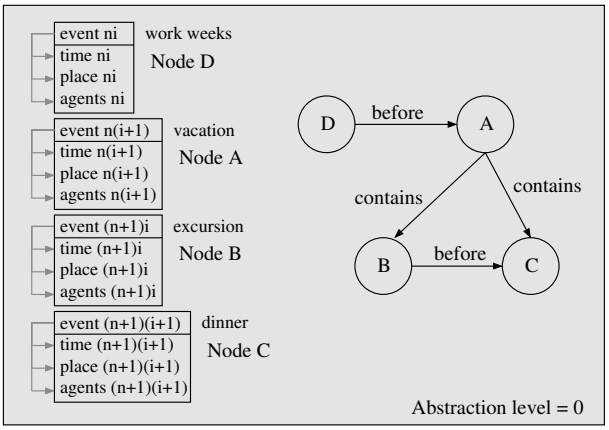

Fig. 6. Semantic description.

"dinner" and `Relation` with the links *contains*, *before*. In this case the relations can be represented using the CS `TemporalRelationCS`.

```
<Mpeg7 ...>
  <Description xsi:type="SemanticDescriptionType">
    <Semantics>
      <AbstractionLevel dimension="0"/>
      <Label/>
      <SemanticBase xsi:type="EventType"
                    id="E1_vac2002"> ...
      <SemanticBase xsi:type="SemanticPlaceType"
                    id="SP1_vac2002"> ...
      <SemanticBase xsi:type="SemanticTimeType"
                    id="ST1_vac2002"> ...
      <SemanticBase xsi:type="AgentObjectType"
                    id="AO1_vac2002"> ...

      <SemanticBase xsi:type="EventType"
                    id="E1_exc2002"> ...
      <SemanticBase xsi:type="SemanticPlaceType"
                    id="SP1_exc2002"> ...
      <SemanticBase xsi:type="SemanticTimeType"
                    id="ST1_exc2002"> ...
      <SemanticBase xsi:type="AgentObjectType"
                    id="AO1_exc2002"> ...

      ...

      <Graph>
        <Node id="Vacation" href="E1_vac2002"/>
        <Node id="Excursion" href="E1_exc2002"/>
        <Node id="Dinner" href="E1_din2002"/>
        <Relation source="#Vacation" target="#Excursion"
             type="urn:mpeg:mpeg7:cs:
                   TemporalRelationCS:2001:contains" />
        <Relation source="#Vacation" target="#Dinner"
             type="urn:mpeg:mpeg7:cs:
                   TemporalRelationCS:2001:contains"/>
        <Relation source="#Excursion" target="#Dinner"
             type="urn:mpeg:mpeg7:cs:
                   TemporalRelationCS:2001:precedes""/>
      </Graph>
    </Semantics>
  </Description>
</Mpeg7>
```

The abstraction level of the obtained description is set to zero because it represent a concrete semantics of an audio–visual document.

6. CONCLUSION AND FUTURE WORKS

In this work we propose an annotation tool that helps the user to generate a description with semantic information he/she wants to associate to audio-visual material. The resulting description is based on the hierarchical structure of events in time. The events represent semantic entities with associated features of time, place and agents objects. In this phase, the proposed approach provides to the user the possibility to create personalized classification schemes (CS) to represent events and relationships between objects. Indeed, the interface records all CS created by the users; in this way, the user can store his/her CSs (for instance, the relationships between relatives), use it all the times he/she wants and eventually extend it (for instance, when in the family a baby is born). In this first phase, the annotation tool works in a semi–automatic way because there is an interaction with the user. Instead, in the second phase, the input information is further processed in order to build a rich semantic description provided by the user at the time of construction of the structural description. The implementation of this step is straightforward because most of the semantic entities are already defined during the interactive generation of the structural description. Thanks to the analysis of the structural decomposition a graph can be constructed.

Concluding, we can say that the proposed interactive tool is characterized by two important proprieties: reusability and extendibility. In addition, a semantic description can be generated automatically from the structural description produced during the annotation phase.

In future works, we will investigate if it is possible to extrapolate other information from the current descriptions provided by the interactive tools, like for instance more abstract concepts associated to the collection of events. We will also investigate other approaches to obtain explicitly or implicitly semantic information from user interaction. In both cases, the semantic abstraction on different levels could be investigated considering other CS, eventually extended to other entities (for instance, concepts). The structure of the CS could also be enriched using a graph instead of a tree structure, which simply defines a hierarchy.

7. ACKNOWLEDGMENTS

The authors would like to thank S. Masneri for the effort in the development of the interactive tool.

8. REFERENCES

[1] http://caliph-emir.sourceforge.net/.

[2] http://iis.joanneum.at/mpeg-7/.

[3] http://ltswww.epfl.ch/∼newuma/.

[4] http://www.alphaworks.ibm.com/tech/videoannex.

[5] http://www.research.ibm.com/videoannex/index.html.

[6] http://www.ricoh.co.jp/src/multimedia/movietool/index.html.

[7] W. Bailer, H. Mayer, H. Neuschmied, W. Haas, M. Lux, and W. Klieber. Content-based video retrieval and summarization using mpeg-7. In *Proc. Internet Imaging*, pages 1–12, San Jose, CA, USA, January 2004.

[8] W. Bailer, H. Mayer, H. Neuschmied, W. Haas, M. Lux, and W. Klieber. *Content-based video retrieval and summarization using MPEG-7*. In *Proc. Internet Imaging V*, pages 1–12, San Jose, CA (USA), January 2004.

[9] Tsuhan Chen. *Low-Level Features to High-Level Semantics: Are We Bridging the Gap?* In *EWIMT*, London, UK, November 2004.

[10] ISO/IEC 15938-5. *Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Schemes*, September 2001.

[11] M. Lux, J. Becker, and H. Krottmaier. Semantic annotation and retrieval of digital photos. In *Forum at the 15th Conference on Advanced Information Systems Engineering (CAiSE 2003)*, pages 85–88, June 2003.

[12] M. Naphade, C.-Y. Lin, J. R. Smith, B. L. Tseng, and S. Basu. Learning to annotate video databases. In *SPIE Electronic Imaging 2002 - Storage and Retrieval for Media Databases*, San Jose, CA, USA, January 2002.

[13] B. L. Tseng, C.-Y. Lin, and J. R. Smith. Video personalization and summarization system. In *SPIE Photonics East 2002 - Internet Multimedia Management Systems*, Boston, MA, USA, August 2002.

[14] B. L. Tseng, C.-Y. Lin, and J. R. Smith. Video summarization and personalization for pervasive mobile devices. In *SPIE Electronic Imaging 2002 - Storage and Retrieval for Media Databases*, San Jose, CA, USA, January 2002.