MULTIMEDIA DOCUMENT DESCRIPTION BY ORDERED HIERARCHIES: THE TOCAI DESCRIPTION SCHEME

N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, Lorenzo A. Rossi

Department of Electronics for Automation - University of Brescia
Via Branze 38, 25123 Brescia – Italy
{adami, bugatti, leon, pier, lrossi}@ing.unibs.it

ABSTRACT

In this work we present the ToCAI (Table of Content-Analytical Index) framework, a Description Scheme (DS) for content description of audio-visual (AV) documents. The idea for such a description scheme comes out from the structures used for indexing technical books (table of content and analytical index). This description scheme provides therefore a hierarchical description of the time sequential structure of a multimedia document (ToC), suitable for browsing, together with an "Analytical Index" (AI) of the key items of the document, suitable for retrieval. The AI allows to represent in a ordered way the items of the AV document which are most relevant from the semantic point of view. The ordering criteria are therefore selected according to the application context. The detailed structure of the DS is presented by means of UML notation as well and an application example is shown.

1. INTRODUCTION

Nowadays more and more AV material arises from a large variety of digital sources. Thus, there is the need to provide frameworks for an efficient navigation or browsing through the large amount of available material and to retrieve relevant information according to user requirements.

For the aforementioned purposes, in the last years, there have been several contributions in the field of multimedia indexing [4], [8]. Furthermore, the International Standard Organization (ISO) started in October 1996 a standardization process for the description of the content of multimedia documents, namely MPEG-7 [6].

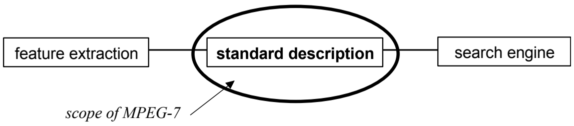

Figure 1. Scope of MPEG-7.

In Figure 1, is shown a block diagram of a possible MPEG-7 processing chain. This chain includes feature extraction (automatic or semi-automatic), the description itself, and the search engine (application). Despite their usefulness, automatic and semi-automatic feature extraction methods are outside the scope of the standard in order to allow industry and scientific competition and to benefit of the expected improvements in these technical areas. For similar motivations, the search engines as well will not be specified within the scope of MPEG-7

This standardization effort should bring by September 2001 the definition of a set of standard Descriptors (D) and Description Schemes (DS) expressed according to a Description scheme Definition Language (DDL).

The DS herein proposed rely on a joint approach that takes into account both audio and video processing for constructing a hierarchical organization of audio-visual information [2],[1]. The proposed DS aims at providing the following functionalities:

- ToC DS: characterize the temporal structure of a multimedia document from a semantic point of view at multiple levels of abstraction, so as to have a series of consecutive segments which are coherent in terms of the semantic of information at that level.
- AI DS: allow an easy way to effectively retrieve relevant information, such as objects appearing in the video (e.g., Bill Clinton), or identify specific events of interest (e.g., a murder in a thriller movie or a goal in football match). To these ends, it is important that the objects or events are arranged in an appropriately designed index, according to criteria meaningful for the application context.
- Context DS: offering general and specific information about the content of the multimedia document such as authors, title, production's date, etc.
- Metadata DS: provide useful information about the document description itself like, e.g., the size of the description and the type of involved extraction methods with a reliability value associated to each automatically extracted descriptor.

The original idea for such a DS originates from the structures adopted to describe information content in technical books. Indeed one is able to easily understand the sequential organization of the book by looking at the table of content while a quick search of elements of interest can be achieved by means of the analytical index of keywords. The ToCAI allows a similar mechanism to address multimedia material in the analytical index, with a couple of extensions: it allows to retrieve information at any given level of abstraction, which is not normally the case in a book (each keyword in the index points normally to the page numbers only, not the sections or paragraphs where the topic of interest can be found). It also allows to arrange the items of the analytical index (key items) according to ordering criteria relevant to the involved application context.

The contribution is organized as follows. In Section 2, the structure of the ToCAI DS is explained by means of UML notation. In Section 3, the issue of the automatic description creation is shortly addressed. Finally some details about the adopted visual interface are given and an example of implementation of such a DS is shown (Section 4).

2. STRUCTURE OF TOCAI DS

We describe the ToCAI structure by presenting the hierarchical organization of its sub-description schemes and involved descriptors. The ToCAI is organized in four main DSs: the *Table of Contents* (ToC), the Analytical Index (AI), the *Context* and the *Meta-descriptors* description schemes.

2.1 ToC DS

The ToC describes the temporal structure of the AV document at multiple levels of abstraction. The lower levels provide a detailed characterization of the temporal structure of the AV document while the higher ones have the role to offer a more compact description with associated semantics.

The ToC is formed by two DSs, namely *Audio-visual Structure* and *Audio Structure*

The Audio-visual structure DS is represented in Figure 2. The two *Time-code Ds* specify the start and the end position of the AV document. The core of this DS is the *Scene DS*. A scene is a temporal segment having a coherent semantics at a certain hierarchical level. It is formed by a various number of sub-scenes, a time reference (2 time-code Ds) and a *type of scene D* (a string and, if useful, a characteristic icon). The elementary component of a scene is the shot[1]. The *Shot DS* indicates the type of editing effects (cut, dissolve, fade in, etc.) and their temporal location (*Editing effects D*). It includes a set of DSs for K-frames mosaic and outlier images of the shot[2]. For example in a soccer match a scene could be a goal, where the two timecodes are the begin and the end of the scene, the sub scenes coulde be the goal, the exultation after that, the replay, the crowd's enthusiasm and so on: at low level each sub scene is composed of shots.

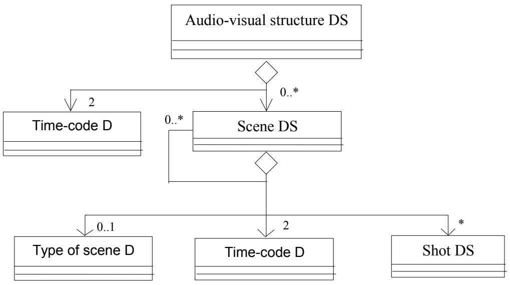

Figure 2. The Audio-Visual structure DS.

The *Audio-structure DS* is organized similarly to the *Audio-visual DS*. In this case, "the leaves of the tree" i.e. audio segments corresponding to a homogeneous audio source are represented by the *Homogeneous audio DS* instead of the *Shot DS*. For more details see [2] or [1].

2.2 Analytical Index DS

The *Analytical Index DS* consists of the ***Key Items DS*** (see Figure 3). The *Key Items DS* can be seen like a generalization of the concept of key words to the context of MM documents. It is composed of several DSs (*Key events DS*, *Key objects DS*, *Key images DS*, *Key sounds DS*), each of them consisting in a set of key items representative for certain type of description element (e.g., events, objects etc.).

In Figure 4, the structure of the *Key Images DS* is shown. The other DSs (*Key events DS*, etc.) are organized in a similar way. It can be seen that each *Key images DS* can contain an arbitrary number of sub-key images, and therefore forms a hierarchy (tree) of key-items. For example, an AV document of a soccer match can contain, at a higher level of this hierarchy (then at a higher level of abstraction), a key image representing *goals*, one representing *penalties*, another one representing *corners*, etc. At a lower level of the hierarchy a goal key image can contains a sets of key frames, each of them representing the shots describing that goal.

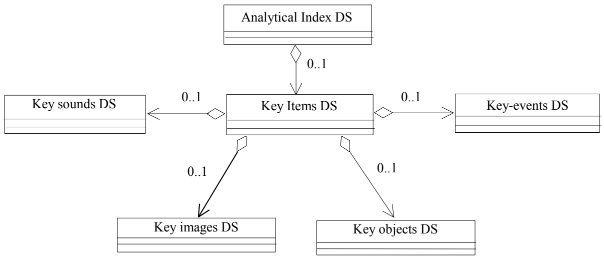

Figure 3. The Key Items DS.

More in detail, the *Key images DS* is formed by the *Ordering Key DS,* the *Attribute DS*, the *Links to segment DS*, and the *Representation DS*. The *Ordering Key DS* aims to order the below *Key images* according to a certain semantic: it presents a list of applicable ordering mechanisms for the listed key items, as you see in Figure 5 (e.g., having defined a set of key-images in a violent movie, the associated ordering key may be the level of underlying audio loudness descriptor and the ratio of red pixel present in the image).

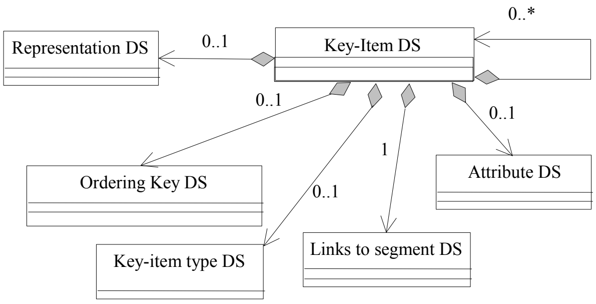

Figure 4: The UML structure of the Ordering Key DS.

According to the tree structure of the *Key images DS*, we can assign different sets of ordering-keys at different key items pertaining at different level of the hierarchy. Obviously these ordering keys can be applied to order other description items.

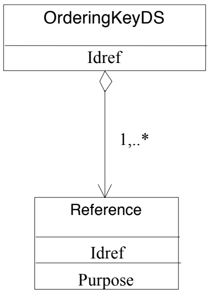

Figure 5: The Key Images DS

The *Attribute DS* characterizes the key item itself. Note

[1] A shot is defined by a sequence of frames captured from a unique and continuous record of camera.

[2] A mosaic represents the background in a shot. An outlier represents a foreground object in motion with respect to the background. These are typically extracted thanks to mosaicing techniques.

that it could be (at least partially) accessed thanks to the *Links to segment DS*. The *Links to segment DS* identifies the parts (e.g., temporal segments, K-frames, …) in the sub-DS the key item refers to. *Links to segment DS* allows clearly to refer to descriptors associated with such parts of the sub-DS. The purpose of the *Representation DS* is the visualization and the presentation of the key-items.

2.3 Meta-descriptors DS

This DS has the role to incorporate in the ToCAI DS a set of descriptors carrying information about how accurate is the description and by which means it has been obtained. The goal is to describe not the content, but to give an indication of the reliability with which the descriptor values have been assigned. Therefore it is of importance to let the user know the identities of the content provider and the description provider (they could be different), the type of involved extraction methods or the size of the description itself.

Besides, a set of descriptors about the reliability level of involved extraction methods it is useful to give users an idea about how much they can trust a given description for answering their queries.

2.4 Context DS

The ToCAI should be considered together with a DS describing the category of the audio-visual material. This context DS includes descriptors such as title of programme, actors, director, language, country of origin, etc. Indeed this kind of information is necessary for retrieving purposes to restrict the search domain, thus facilitating the retrieval performance of a query engine [2].

3. EXTRACTION METHODS

We have adopted several tools in order to obtain automatic feature extraction for describing an AV document according to ToCAI structure.

Individual shot separation is achieved by extraction of editing effects between consecutive camera records. This can be obtained by making use of the statistical independence of the two shots that are present on both sides of the editing effect; in the case of dissolves, fade-in, or fade-our, refer to the algorithm presented in [3].

Shot grouping into scenes is obtained by identification of peculiar alternation of visual patterns between consecutive shots, so as to recognize characteristics situations such as dialogues, actions, … The visual correlation between non consecutive shots is established thanks to a vector quantization approach, which compares codebooks associated to the individual shot patterns [10].

On the side of the Key Items extraction, it is necessary to focus on the involved application context for developing ad-hoc automatic extraction tools. We developed and tested some simple automatic extraction methods to order shot objects in our test material. For example, in a football match video, we estimated the average audio loudness related to each shots. The shots having the higher audio loudness turned out to be very meaningful from the semantic point of view (goal, roar crowd after goals, …). Moreover in a news programme the dominant hue was allowed to cluster the shot associated with the news speaker. Another feature used was a measure of the pan camera motion, which has been very useful in sport documents to extract the shots with a camera closed to a player having fast movement. For further details about the experiment results see [11].

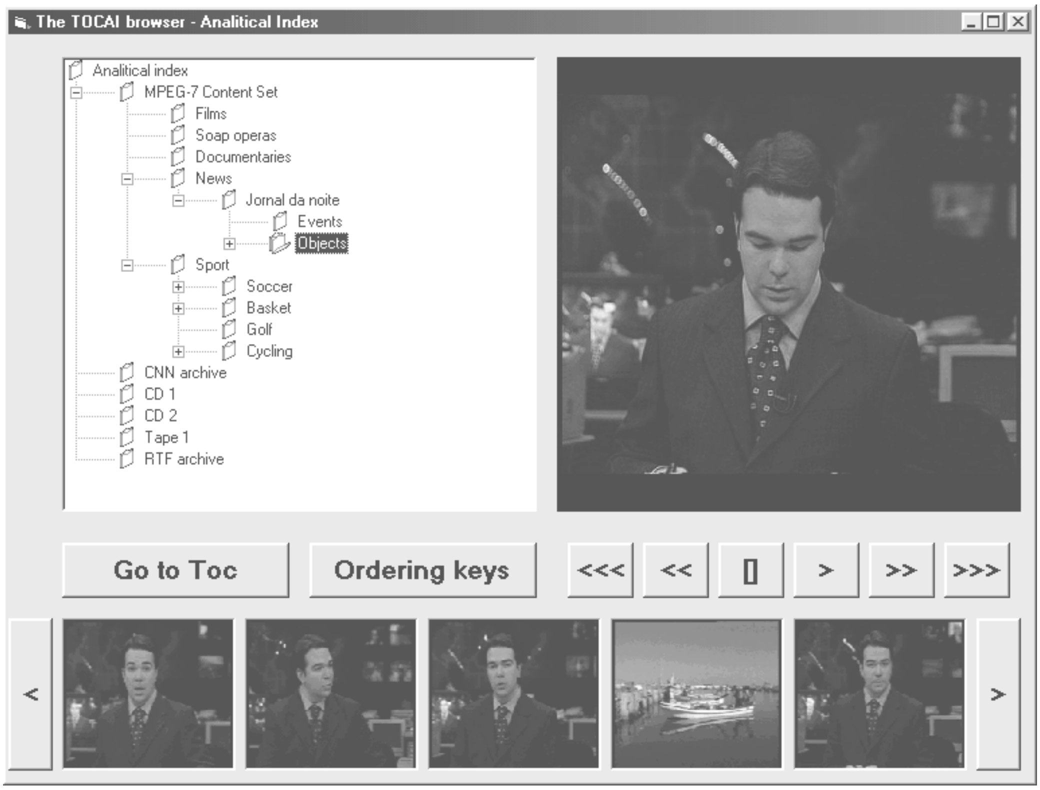

Figure 6. The visual interface for the Analytical Index DS

4. APPLICATION EXAMPLE

For exploiting a ToCAI description of a multimedia document, it is necessary to adopt a visual interface. For such a purpose, we have developed a Visual Basic application. In Figure 6, it is represented the visual interface for the Analytical Index DS. It is very similar to the one of the Table of Content, since in both cases a certain kind of ordered structure is represented (in the latter case the order is chronological).

The ToCAI DS can be adopted for several browsing and retrieval tasks such as rapid navigation throughout multimedia material (at the scene and/or shot level), retrieval of objects or events of interests etc. This example show how the Analytical Index DS could be applied to a set of programme archives of multimedia documents. In this particular case, the Key items DS recursively points to other Key Items DSs associated to several archives and to certain categories of programs (Sports, News, etc.).

In Figure 6, is represented the hierarchical structure for the key items (left part). Each key item has an associated Ordering Key DS, which is a list of useful ordering keys for the key items belonging to the below level. For example, the key item Sport has three associated ordering keys: the name, the personal preferences and a sort of index of the dynamism. In particular, a set of shots belonging to a news programme is ordered according to their level of the dominant hue. How it can be seen in the image containers at the bottom of the form, the order induct to this Ordering Key allows to browse the shots containing the speaker figure, despite their temporal position. The interface allows to play the corresponding sequences as well as to select the involved scene (or shot) of the ToC DS, for chronologically navigating throughout the multimedia material.

5. CONCLUSION

The paper presented the ToCAI DS as a framework for multimedia content description, which provides nice navigation and retrieval functionalities. The proposed audio-visual DS is based on four main structures:

-A *Table of Contents DS* for semantically characterizing the temporal structure of the multimedia document.

-An *Analytical Index DS* for providing an ordered set of relevant objects of the document with links to the document itself.

-A *Context DS* for focusing on the category of the document.

-A *Meta-descriptors DS* for giving useful information about the description itself and its reliability.

The detailed structure of the DS has also been presented, and an application example for navigation and retrieval was shown.

Current research is devoted to the study of suitable automatic extraction methods, so as to generate the different D's which are part of the ToCAI DS in an automatic way. Another research effort is also being carried out to identify the extension which should be added to the generic AV DS currently under study by ISO/MPEG for the MPEG-7 standard [7].

6. REFERENCES

[1] N. Adami, A. Bugatti, A. Corghi, R. Leonardi, P. Migliorati, L. A. Rossi, C. Saraceno, "ToCAI: a framework for Indexing and Retrieval of Multimedia Documents", *In Proc. International Conference on Image Analysis and Processing (ICIAP'99)*, Venice, Italy, Sept. 1999.

[2] R. Leonardi *et al.,* "The ToCAI description scheme for indexing and retrieval of multimedia documents", *In Proc. European Workshop on Content-Based Multimedia Indexing (CBMI'99)*, Toulouse, France, Oct. 1999.

[3] N. Adami and R. Leonardi, "Identification of editing effects in image sequences by statistical modeling", In *Proc. Picture Coding Symposium '99*, Portland, OR, U.S.A., Apr. 1999.

[4] A. Ferman, A. Tekalp and R. Mehrotra, "Effective content representation for video", In *Proc. IEEE International Conference Image Processing*, Chicago, IL, Oct. 1998.

[5] M. Fowler, *UML Distilled*, Addison Wesley, Longman, 1997.

[6] MPEG-7 Requirement Group, "MPEG-7: Context and objectives", *ISO/IEC JTC1/SC29/WG11* N2460, MPEG98, Atlantic City, USA, Oct. 1998.

[7] MPEG Description Scheme Group, "MPEG-7 Description Schemes (V0.5)", *ISO/IEC JTC1/SC29/WG11* N2844 MPEG99 Vancouver, Jul. 1999.

[8] Y. Rui, T. Huang and S. Mehrotra, "Browsing and retrieving video content in a unified framework", In *Proc. IEEE Workshop on Multimedia Signal Processing,* Dec. 1998.

[9] C. Saraceno and R. Leonardi, "Indexing audio-visual databases through a joint audio and video processing", *International Journal of Imaging Systems and Technology*, 9(5):320-331, Oct. 1998.

[10] C. Saraceno and R. Leonardi, "Identification of story units in audio-visual sequences by joint audio and video processing", In *Proc. International Conference on Image Processing 1998*, Chicago, IL, U.S.A., Oct. 1998.

[11] N. Adami, A. Bugatti, R. Leonardi and L. Rossi, "Validation Experiment on the Ordering Key DS and an Unified Syntax for the Weight DS" *ISO/IEC JTC1/SC29/WG11* M5605, MPEG99, Maui, USA, Dec. 1999.