The ToCAI Description Scheme for Indexing and Retrieval of Multimedia Documents[1]

N. Adami, A. Bugatti, A. Corghi, R. Leonardi, P. Migliorati, Lorenzo A. Rossi, C. Saraceno[2]

Department of Electronics for Automation
University of Brescia
Via Branze 38, I-25123 Brescia – Italy
PHONE: +39-030-3715434. FAX: +39-030-380014
Email: leon@ing.unibs.it

Abstract. In this work, a framework, called Table of Content-Analytical Index (ToCAI), for the content description of multimedia material is presented. The idea for such a description scheme (DS) originates 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 audio-visual objects of the document, suitable for effective retrieval. Besides other two sub-description schemes for the category and the description of the meta-data associated to the multimedia document are enclosed in the general DS. The detailed structure of the DS is presented by means of UML notation as well, and an application example is shown.

1 Introduction

Nowadays a huge amount of audio-visual (AV) material arises from a variety of digital sources. Therefore there is the need of suitable frameworks for efficient browsing through the available material and for retrieving relevant information according to user requirements.

For the aforementioned purposes, in the last years, there have been several contributions in the field multimedia indexing [2][6][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 [4][5]. 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). A DS can be used to generate a description of a multimedia document with various levels of abstraction, by combining descriptors characterizing features such as shape, color, texture, motion (for the video component), or audio type (for the audio component) [1]. The DDL should allow to build a variety of different new description schemes for dealing with specific application contexts.

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.

The proposed DS aims at providing the following functionalities.

- 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. With this kind of indexing procedure, a fast navigation throughout a multimedia document can be carried out.
- Allow an easy way to effectively retrieve relevant information, such as objects appearing in the video (e.g., Bill Clinton), or identify specific situations of interest (e.g., a murder in a thriller movie or a goal in football match). To have a good retrieval capability, it is important that these objects or events be arranged in an appropriately designed index, according to various criteria, so as to ease the retrieval task.
- Offering general and specific informations about the content of the multimedia document such as authors, title, production's date, etc.
- Provide useful informations about the document description itself like, e.g., the size of the description and the type of involved extraction methods with a confidence interval associated to each descriptor value (a "reliability" descriptor).

[1] This work has been partially funded by the European ESPRIT project AVIR (Audio-Visual Indexing and retrieval for non IT expert users).

[2] Caterina Saraceno is currently with Starlab NV, Excelsiorlaan 40, B-1930 Zaventem, Belgium.

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, generally placed at the end of the book. In the first case, the chronological order of presentation is preserved, while in the last case, an alphabetical order exists to facilitate the retrieval task. 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 elements of the analytical index in various manners (not necessarily an alphabetical order of keywords), according for example to the individual user preferences.

This paper is organized as follows. Section 2 presents the main functionalities of the ToCAI DS and gives a detailed explanation of the its structure with the involved sub-DSs and Ds, using the Universal Modeling Language (UML) notation [3]. In Section 3, an example of implementation of such a DS is shown, with a an example of possible browsing and retrieval functionalities. Finally section 4 summarizes the presentation and suggests further elements of study to map the Table of Content DS into the current generic AV DS of MPEG-7, and suggests that extensions should be added to such DS to cope with the ordering functionality which is built in the structure of the Analytical Index DS.

2 Structure of the ToCAI DS

We describe now 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. (see Figure 1).

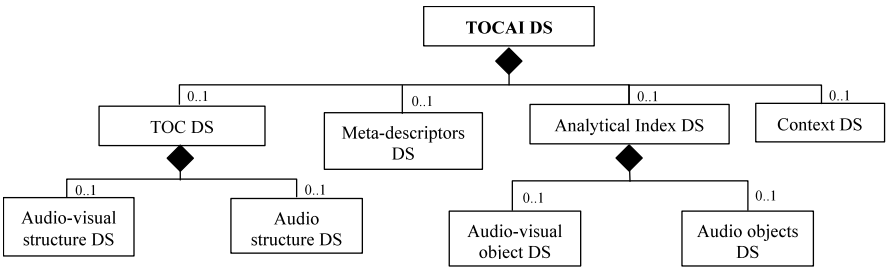

Fig. 1. High level structure of the ToCAI DS.

2.1 ToC DS

The ToC describes the temporal structure of the AV document at multiple level of abstraction. It is organized in different hierarchical levels where the lower levels provide a detailed characterization of the sequential structure of the AV document while the higher ones have the role to offer a more compact description with associated semantics. A key aspect is that the items at each level are kept in **chronological order**.

The ToC DS is very useful for browsing and navigation, since it provides summaries of the document at several levels of details. Besides the meaningful characterization of the temporal structure of the document, elements of the ToC DS may also be used for retrieval tasks, by restricting the search field for a particular query.

The ToC is formed by two DSs described hereafter, namely *Audio-visual Structure* and *Audio Structure*. We proposed not to identify a simple visual DS as it is in general difficult or meaningless to temporally decompose at a high level of abstraction, a sequence of images without making use of the associated audio information [7].

Audio-visual structure DS

This 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[3].

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

The *Shot DS* indicates the type (cut, dissolve, fade in, etc.) of editing effects and their temporal location (*Editing effects D*). It includes a set of DSs for K-frames mosaic and outlier images of the shot[4].

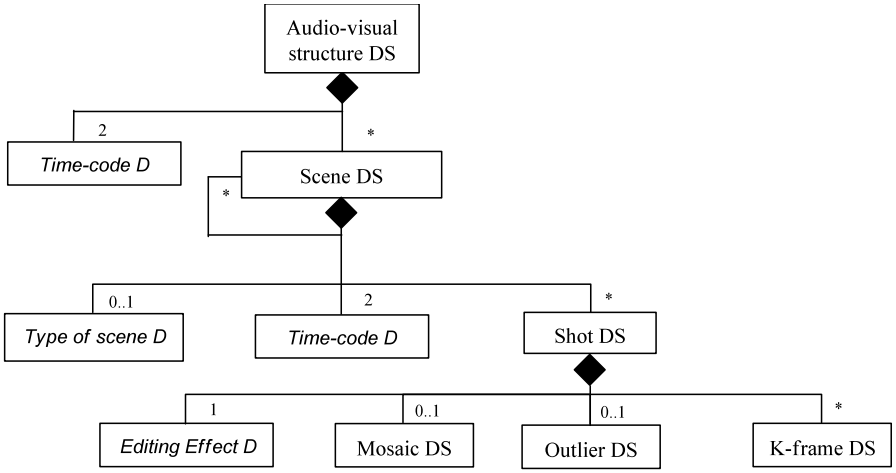

Fig. 2. The Audio-visual structure DS.

Audio-structure DS

This DS has a similar structure to the Audio-visual DS. Thus we can have various layers of audio scene and herein the *Shot DS* is replaced by the *Homogeneous audio DS*. The Ds associated to the *Homogeneous audio DS* represent the leaves of the tree, i.e. audio segments corresponding to a homogeneous audio source (for example a particular speaker, a particular noise, a defined music etc.). Each homogeneous audio source may be represented in terms of an appropriate label and a time reference.

2.2 Analytical Index DS

The AI allows to create an **ordered** set of audio-visual objects of the multimedia document. An item in the AI can point at different locations and at different levels of abstraction (according to the hierarchy provided in the ToC). Hence this DS has the main role to support retrieval of selected objects within the AV document. It is formed by two DSs: the *Audio-visual object DS* and the *Audio object DS*.

AI objects can be semantic entities (like an AV scene belonging to a particular category, e.g. a dialogue), particular kind of images (backgrounds, foreground objects, etc.) but audio objects as well (like the musical motif and/or some keywords from a speech to text transcription). These objects can be ordered according to various criteria, called ordering keys, which are defined at the time of the index creation and are part of it. It must be pointed out that thanks to the AI, more than one shot or more than one scene may be referenced by the same AI item.

Audio-visual objects DS

The structure of this DS is shown in Figure 3. The *ordering keys D* is set of possible keys for ordering the AI items, e.g. color or texture for images. Two classes of AV objects are foreseen: scenes and images. Consequently there are two main DS's.

For each object of the index, there are several pointer types, which define the level in the ToC hierarchy for which a reference link has been generated. Consequently, for every type of pointer (i.e. level of the ToC hierarchy), several reference pointers are listed.

[4] 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, which allow to register regions at different layers moving differently throughout the image sequence.

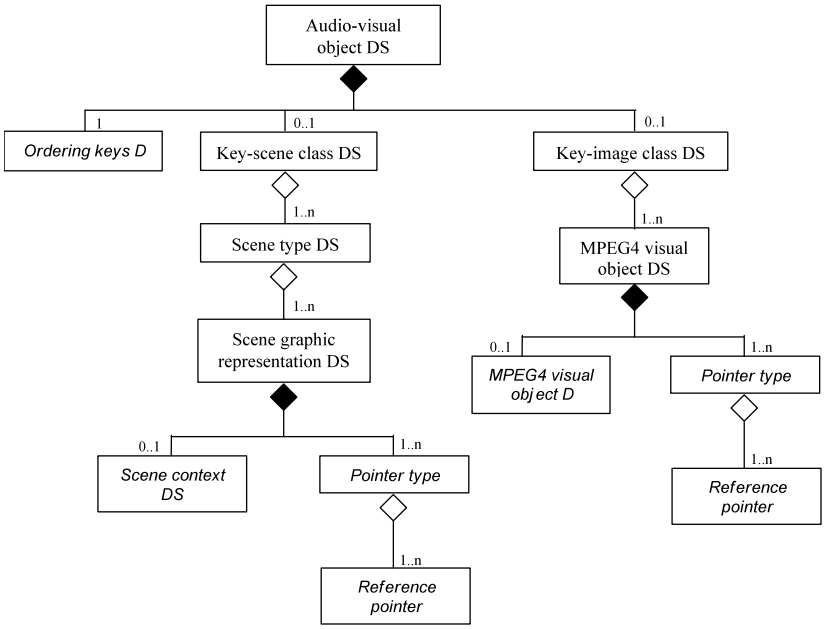

Fig. 3. The Audio-visual object DS.

Audio- objects DS
The structure of this DS is analogous to the one AV object DS. In this case the considered objects may be keywords (provided for example by a speech to text transcription), the identities of the speakers involved in the multimedia document, relevant musical motifs etc.

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 descriptor values have been assigned throughout the ToCAI DS. First, it is of importance to let the user know who are the content provider and the description provider (they could be different). Other important information should consist in the type of involved extraction methods or in the size of the description itself.

Besides, a set of descriptors about the reliability level of involved extraction methods may give users an idea about how much they can trust a given description for answering their query. Thus these descriptors provide a very important complement to the content description itself, since a description generated by an unreliable extraction method may be of little usage for retrieval purposes.

2.4 Context DS

The ToCAI, which refers to the structure of an AV document, should be considered together with a DS describing the category of the audio-visual material. This contextual DS includes descriptors such as title of programme, actors, director, language, country of origin, etc. Indeed these informations are necessary for retrieving purposes to restrict the search domain, thus facilitating the retrieval performance of a query engine.

This DS contains a set of the typical programme descriptors that are readily available, normally at the time the programme was created or archived. In a TV programme, these correspond to the title of the programme, the country of origin, the year of production etc.

3 Application Example

The ToCAI DS seems very adequate to describe the content of a large AV programme such as a movie. The ToC allows to navigate at different levels of details (scene or shot), while the AI gives the possibility to retrieve individuals or specific events present in the movie.

We show now an implementation example of the ToCAI DS generated to help navigate through and retrieve relevant information in a broadcast news (drawn from the MPEG-7 Content Set). The programme was segmented in shots which have then been clustered in scenes. Every ToC item is represented by a K-frame. In Figure 4, a subset of scenes can be seen. The icons below the K-frames of the scene identify the type of scene (in a TV news programme there are mainly two types of scene: speaker presentations and reportages). By going down one level in the hierarchy, it is possible to identify the individual shots forming the scene and view their associated mosaics. Figure 5 represents an implementation of the AI. A set of shot backgrounds (mosaics) is shown. They have been ordered according to dominant color information and they point to the corresponding scenes, in the ToC. These may be accessed by hitting the "down level" button shown on the display.

It should be also noted that a playback mode functionality if made readily available to view particular portions of the audio-visual programme (at any given scene or shot level).

Fig. 4. Implementation of the ToC.

4 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: 1) a *Table of Contents DS* for semantically characterizing the temporal structure of the multimedia document. 2) An *Analytical Index DS* for providing an ordered set of relevant objects of the document with links to the document itself. 3) A *Context DS* for focusing on the category of the document. 4) 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 extention which should be added to the generic AV DS currently under study by ISO/MPEG for the MPEG-7 standard [9].

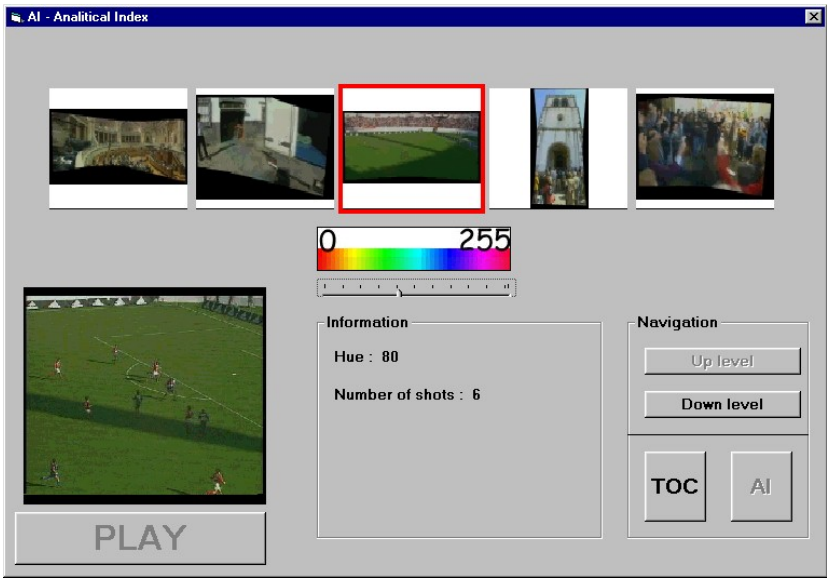

Fig. 5. Implementation of the AI.

References

[1] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, L. Rossi: ISO/IEC JTC1/SC29/WG11/M4586: The TOCAI DS for audio-visual documents. Structure and concepts, Document submitted to MPEG-7, Seoul, Korea, Mar. 1999.

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

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

[4] MPEG Requirement Group. MPEG-7: Context and objective. ISO/IEC JTC1/SC29/WG11 N2460, *MPEG98*, Atlantic City, USA, October 1998.

[5] MPEG Requirement Group. MPEG-7: Requirements. ISO/IEC JTC1/SC29/WG11 N2461, *MPEG98*, Atlantic City, USA, October 1998.

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

[7] C. Saraceno, 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.

[8] S. Smoliar and L. Wilcox. Indexing the content of multimedia documents. *Proc. Second International Conference on Visual Information Systems*, San Diego, CA, 1997.

[9] MPEG-7 Description Schemes (V0.5), MPEG description scheme group, ISO/IEC JTC1/SC29/WG11, MPEG document N2844, Vancouver, Jul. 1999.