ToCAI: a Framework for Indexing and Retrieval of Multimedia Documents

N. Adami, A. Bugatti, A. Corghi, R. Leonardi, P. Migliorati, Lorenzo A. Rossi and C. Saraceno
University of Brescia, Department of Electronics for Automation
Via Branze 38, I-25123 Brescia – Italy
{adami, leon, pier, lrossi}@ing.unibs.it

Abstract

This paper presents the ToCAI (Table of Content-Analytical Index, ToCAI) Description Scheme (DS) for content description of audio-visual documents. The original idea comes out from the structure used for technical books. Ones may easily understand a book sequential organization by looking at its table of contents while quickly retrieve elements of interest by means of the analytical index. This description scheme provides therefore a hierarchical description of the time sequential structure of a multimedia document (thanks to the ToC), suitable for browsing, together with an "Analytical Index" (AI) of audio-visual objects of the document, suitable for effective retrieval. Besides, two sub-description schemes for information about description generation and about the metadata associated to the document are also enclosed in the general DS. The detailed structure of the DS is also presented by means of UML notation and an application example is shown. Finally, some considerations about the adopted visual interface are made.

1. Introduction

Nowadays more and more AV material arises from a variety of digital sources. There is therefore the need to provide frameworks for an efficient navigation or browsing through the large amount of material being made available and to retrieve relevant information it contains according to a specific user.

To help users in the localization of this information, the International Standard Organization (ISO) started in October 1996 a standardization process for the description of the content of multimedia documents, namely MPEG-7: the "Multimedia Content Description Interface" [6],[7]. This standardization effort should bring by September 2001 the definition of a set of standard descriptors (D) and description schemes (DS) expressed by means of a description 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, colour, texture, motion (for objects), or audio type (for the audio component) [1]. The DDL should allow to build a variety of different description schemes which are adequate to deal with a specific application context.

For the aforementioned reasons as well, in the last years, there have been several contributions in the field multimedia indexing and retrieval. See, e.g., [3],[8],[11].

In a description scheme which aims to organize the multimedia information in a hierarchical manner, it is possible to have, at the lowest abstraction level, a description of the content which is limited to shape, size, color and texture of objects forming each individual frame [5]. At a higher abstraction level, the content description can be based on motion characteristics of individual objects together with a description about the deformations that these objects can undergo.

A further abstraction level should be located in correspondence with changes of camera records with the indication of editing effects (e.g., cut, dissolve...) in proximity of each shot transition [2].

If the abstraction process continues, the description becomes progressively more semantic, with the identification of scenes. These can then be interpreted, by adequate knowledge models, if a symbolic description is available (e.g., scene containing a dog that barks to the left, a blue ball that falls to the right, with a background engine noise in the surround).

To reach a high level organization of information, it is essential to give audio and video signals the same level of importance. Unfortunately, research efforts on audio and video processing have been traditionally carried out independently. Only recently, these two sources of information have been jointly considered [9],[10]. From the first simulations, it has been demonstrated that a joint audio and video analysis is effective for the identification, e. g., of simple scenes that compose a multimedia program [10]. For instance, consider an automatic procedure which should detect a salient moment in a soccer game. An algorithm which analyzes the video frames in order to recognize important events, such as the occurrence of a goal, should be combined with an algorithm that tries to identify a rapid increase of the

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

audio energy, which could be associated with a crowd roar.

From this analysis, it is clear that there is a strong interest in the solution of the problems previously described. On the other hand, it is clear that there is the need for joint audio and video processing to achieve a better interpretation.

In this work, we propose a description scheme for the hierarchical description of multimedia documents, by taking into account both audio and video analysis. 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 level of abstraction, so as to have a series of consecutive segments which are coherent with the semantic of information at that level. This allows to reach a global summary for the document, with a possibility to enter detailed levels of description if required. With this type of indexing procedure, a rapid navigation through the 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 killing in a thriller movie). To have a good retrieval capability, it is important that these objects or events be arranged in the AI 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 their reliability factor.

The original idea for such a DS comes out from the structure used for technical books. One may easily understand a book sequential organization by looking at its table of content (generally located in the first pages) and/or may quickly retrieve elements of interest by means of the analytical index (typically located 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. The ToCAI allows a similar mechanism to address multimedia material, with one extension: 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).

The paper is organized as follows. Section 2 gives some quick concepts about MPEG-7 [6]. Section 3 introduces the structure of the ToCAI DS presenting the main functionalities of its sub-DS. Section 4 gives a detailed explanation of the ToCAI structure with its involved DSs and Ds using UML notation. In Section 5, an example of implementation of such a DS is shown and in Section 6, some technical details about the visual interface design are given.

2. MPEG-7 context and objectives

In October 1996, MPEG started a new work item to address the issue of the multimedia content description: the "Multimedia Content Description Interface" (in short "MPEG-7"). The purpose of MPEG-7 is the specifica-tion of a standard set of descriptors that can be used to describe several kinds of multimedia information. MPEG-7 will also standardize structures (Description Schemes) for the descriptors and their relationships as well as a language for specifying description schemes, i.e. a Description Definition Language (DDL). The standard will be applicable to AV material like still pictures, graphics, 3D models, audio, speech, video.

An MPEG-7 description may be either physically located with the associated AV material or also live somewhere else on the globe. When the content and its descriptions are not co-located, mechanisms that link AV material and their MPEG-7 descriptions are useful; these links should work in both directions.

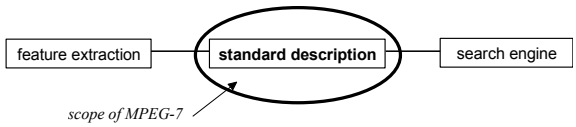

Figure 1. Scope of MPEG-7.

In Figure 1, is shown a block diagram of a possible MPEG-7 processing chain. This chain includes feature extraction (analysis), the description itself, and the search engine (application). Automatic extraction of features (or 'descriptors') will be extremely useful for a full exploitation of MPEG-7 descriptions. However automatic extraction is not always possible (the higher the level of abstraction, the more difficult automatic extraction is).

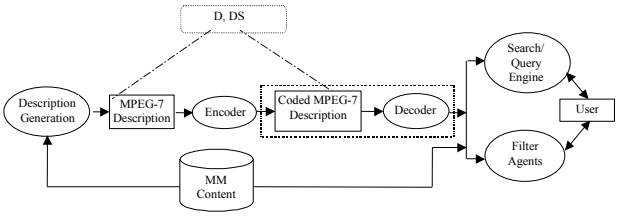

Figure 2. An abstract representation of possible applications using MPEG-7.

Hence interactive extraction tools will be of good use. Despite their usefulness, automatic and semi-automatic feature extraction methods are outside the scope of the standard since their standardization is not required to allow interoperability, while leaving space for industry competition. Another reason not to standardize analysis is to allow making good use of the expected improvements in these technical areas. Due to the previous motivations, the search engines as well will not be specified within the

scope of MPEG-7.

Figure 2 explains how MPEG-7 would work in practice. The emphasis of MPEG-7 will be the provision of novel solutions for audio-visual content description. More detailed explanation about MPEG-7 can be found in the documents [6],[7].

3. ToCAI overview

This DS is created by the aggregation of four main description schemes: the Table of Contents (ToC), the Analytical Index (AI), the Context and the Meta-descriptors description schemes.

ToC DS
The ToC DS 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.

For example, in a broadcast news document, we can have a ToC description like the following:

- Summary (0:2'30'')
- Internal affairs (2'31':7')
  - Speaker presentation (…)
  - First reportage (…)
    - Shot 1 (…)
    - Shot 2 (…)
    - …
  - Speaker presentation (…)
  - ….
- International affairs (…)
  - ….

Where the data within parenthesis specify the temporal location of the segment, while the label indicates its semantics. Every ToC item may be summarized by key-frames and audio segments.

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 characterisation of the temporal structure of the document, provided by the ToC DS, may also be used for retrieval tasks as it can restrict the search field for a particular query, given the hierarchical structure which is created. As a summary two words are essential in the ToC concept:

- hierarchy,
- chronological order.

Analytical Index DS
The AI allows to create an **ordered** set of audio-visual objects. An item in the AI can point at different levels of detail according to the hierarchy provided in the ToC, or according to some other criteria. It must be pointed out that thanks to the AI, more than one shot or more than one scene can be referenced by the same AI item. This allows to navigate along the audio-visual material not in sequential order, rather through scenes/shots containing similar objects.

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, which are listed in the DS. As a summary two words are essential in the AI concept:

- order,
- reference pointer.

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.

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 purpose is to describe not the content but to give an indication about the reliability of the descriptor value assignment.

4. Detailed structure of the proposed DS

We describe now the ToCAI structure by presenting the hierarchical organization of its sub-description schemes and involved descriptors (see Figure 3). We adopted the Universal Modeling Language (UML) notation [4].

4.1 ToC DS

It describes the temporal structure of the AV document at multiple level of abstraction. It contains two DSs, explained below, namely *Audio-visual Structure* and *Audio Structure*. We proposed an *Audio-Visual Structure* DS rather than a simple visual DS because, from the semantic point of view, it is often necessary to consider the information carried by video together with the one provided by associated audio so that to recover reliable intermediate semantic levels for the description.

Audio-visual structure DS
This DS is represented in Figure 4. 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 semantic at a certain hierarchical level. It is composed by a various number of sub-scenes, a time reference (two 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,

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

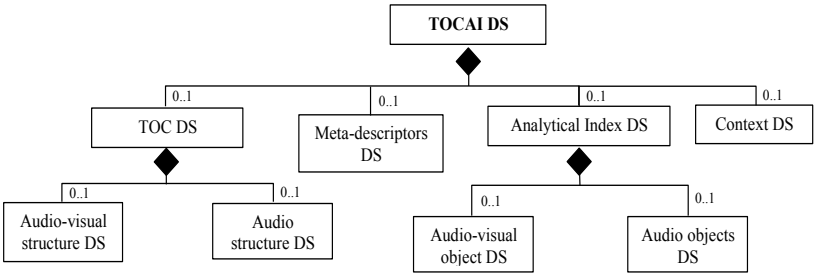

Figure 3. High level structure of the ToCAI DS.

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]

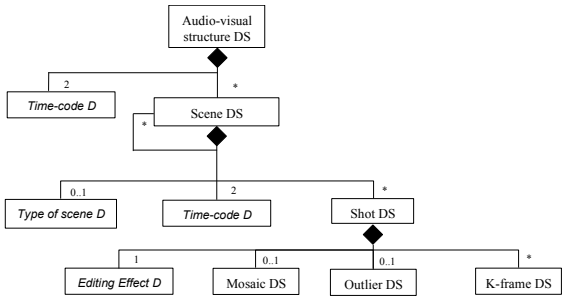

Figure 4. The Audio-visual structure DS.

Audio-structure DS
This DS reflects the structure of Audio-visual DS. Hence we can have various layers of audio scene. The Ds belonging to the *Homogeneous audio DS* represent the leaves of the tree, i.e. audio segments having an homogeneous audio source (for example a particular speaker, a particular noise, a defined music etc.). Each one of these Ds is constituted by an appropriate label and a time reference.

4.2 Analytical Index DS

As we said, the AI allows to create an ordered set of audio-visual objects of the document pointing at different locations and different level of abstraction. Therefore this DS has the main role to support retrieval of selected objects within the AV document. It consists of two DSs: the *Audio-visual object DS* and the *Audio object DS*.

Audio-visual objects DS
The structure of this DS is shown in Figure 5. The *ordering keys D* is set of possible keys for the ordering of AI items, e.g. colour or texture for images. Two classes of AV objects are foreseen: scenes and images. Consequently there are two main DSs.

[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.

For each object of the index, there are several types of pointers (an object may point at different levels of abstraction) and for every type of pointer, several reference pointers.

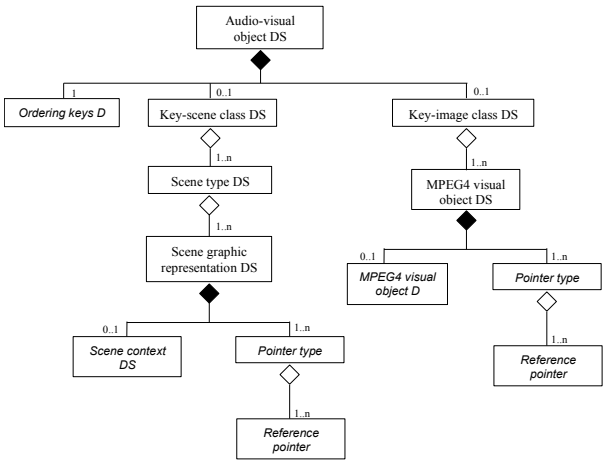

Figure 5. The Audio-visual object DS.

Audio- objects DS
This DS is similar to the audio-visual object DS. The *ordering keys D* is set of possible key for the ordering of AI items, e.g. name of musical instruments, time duration of an audio segment.

4.3 Meta-descriptors DS

This DS consists of descriptors describing the other descriptors. These meta-descriptors should be chosen so that they can provide indirect but useful information about the content of a multimedia document. First, it is of importance to let the user know the identities of the content provider and the description provider (they could not be the same entity). Other relevant information should consist in the type of involved extraction methods and in the size of the description itself (see Figure 6).

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 query result. Hence these descriptors can be a very important complement to the content description itself, since, e.g., a

description achieved by means of a quite unreliable extraction method should not be very helpful for the content understanding of a multimedia document.

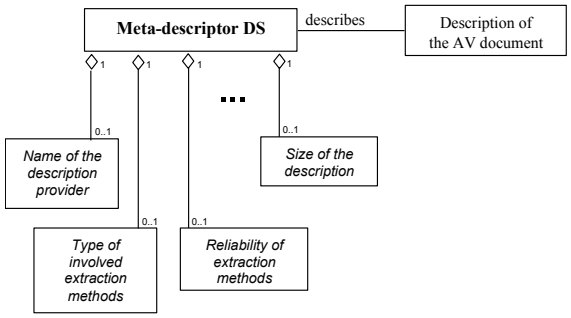

Figure 6. Structure of the Meta-descriptors DS.

4.4 Context DS

This DS consists of a set of typical programme descriptors that are available, e.g., in a Radio-TV programme guide (see Figure 7) like, e.g. the title of the programme, the country of origin, the year of production etc. For example, by means of the knowledge of a movie director, a particular search can be facilitated.

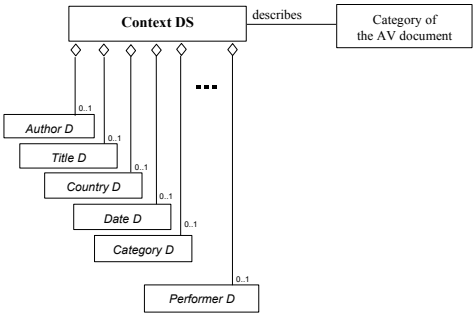

Figure 7: *Structure of the Context DS*

5. 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 backgrounds present in the movie. The proposed DS can, e.g., satisfy queries like the followings:

- "I want to have a quick (and/or more detailed) idea about the content of such AV document viewing a frame (and/or listening to an audio sample) for each of the most representative scenes".
- "I want to see an ordered list of the main objects of the AV document and select the scenes where they occur".
- "How much I can trust a certain query?".

In this section, we show an implementation example of the ToCAI DS applied to a broadcast news programme (MPEG-7 Content Set). The programme was segmented in shots then clustered in scenes. Every ToC item is represented by a key-frame. In Figure 8, a subset of scenes can be seen. The icons below the key-frames representing scenes identify the type of scene (in a TV news programme there are mainly two types of scene: speaker presentations and reportages).

6. The visual interface

The aim of this interface is to allow the browsing in multimedia document by means of description, organized according to the ToCAI DS, in a natural and intuitive way. According to this, we have developed a Visual Basic simple application. To avoid confusion and to let the users focus only on the conceptual part, the number of active objects is kept low.

As we said, the ToCAI consists of two main complementary parts, the Table of Contents and the Analytical Index (Figure 9), which present some strong analogies. They represent a kind of order, in one case a temporal order and in the other case an order inducted by several possible criteria (e.g., the value of hue of an image). To show this characteristic, both the parts are represented in a similar manner.

Every part is composed by several levels and the same buttons are used to navigate through them. When the user find an interesting part of the video stream, he/her can play it on the left bottom of the form. According to the level, a different object is shown.

7. Conclusion

This paper presents the ToCAI description scheme as a framework for the multimedia content description. The proposed DS is based on four main structures. 1) A *Table of Contents DS* for semantically characterizing the temporal structure of an AV document; 2) *an Analytical Index DS* for providing an ordered set of relevant objects of the document with their link to the document itself; 3) a *Context DS* for focusing on the category of the programme; 4) a *Meta-descriptors DS* for providing users with information about the description itself and its reliability. The detailed structure of the DS has also been presented and an application example was shown.

Current research work is devoted to developing suitable automatic extraction methods for the computing the values of the involved descriptors.

References

[1] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati and L. Rossi. The TOCAI DS for audio-visual documents. Structure and concepts. *ISO/IEC JTC1/SC29/WG11/* **M4586**, MPEG99, Seoul, Korea, Mar. 1999.

Figure 8. Implementation of the ToC. Some key-frames linked to scenes of the programme are shown.

Figure 9. The AI. The five mosaics shown are ordered according to a color criterion (hue level).

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

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

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

[5] O. N. Gerek, and Y. Altunbasak: Key Frame Selection from MPEG Video Data. In *Proc. SPIE Visual Communications and Image Processing*, 3024:920-925, 1997.

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

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

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