The DICEMAN Description Schemes for Still Images and Video Sequences

N. O'Connor[1], P. Salembier[2], P. Correia[3], L. Ward[1]

[1]Dublin City University, Dublin, Ireland. Email: {oconnorn,wardl}@teltec.dcu.ie

[2]Universitat Politècnica de Catalunya, Barcelona, Spain. Email: philippe@gps.tsc.upc.es

[3]Instituto Superior Técnico, Lisboa, Portugal. Email: paulo.correia@lx.it.pt

Abstract

To address the problem of visual content description, two Description Schemes (DSs) developed within the context of a European ACTS project known as DICEMAN, are presented. The DSs, designed based on an analogy with well-known tools for document description, describe both the structure and semantics of still images and video sequences. The overall structure of both DSs including the various sub-DSs and descriptors (Ds) of which they are composed is described. In each case, the hierarchical sub-DS for describing structure can be constructed using automatic (or semi-automatic) image/video analysis tools. The hierarchical sub-DSs for describing the semantics, however, are constructed by a user. The integration of the two DSs into a video indexing application currently under development in DICEMAN is also briefly described.

I Introduction

With the increasing availability of digital audio-visual (AV) content (e.g. via the Internet) it is becoming increasingly difficult for users to accurately locate and retrieve content which is particularly suited to their requirements. The ISO/IEC Content Description Interface initiative (more commonly referred to as MPEG-7) aims to standardise a way of describing AV content in order to facilitate flexible and efficient content-based user queries [1][2]. In the context of MPEG-7, a description of an AV document includes Descriptors (termed Ds), which specify the syntax and semantics of a representation entity for a feature of the AV data, and Description Schemes (termed DSs) which specify the structure and semantics of a set of Ds and DSs [3]. Descriptions are expressed in a common description definition language (DDL) to allow their exchange and access. This paper presents two DSs for visual content, which were developed within the framework of a European ACTS project known as DICEMAN[1].

II Overview of the DICEMAN project

Currently, content holders have vast quantities of AV material they would like to sell whereas content consumers have a growing need to access and purchase this content. However, today's business model for content exchange is primarily based on paper, videotapes and CDs, and the postal system, making it slow, expensive and cumbersome. The ACTS DICEMAN project intends to demonstrate an enhanced content exchange business model which opens up databases of multimedia content by making them easier to search over electronic networks using advanced forms of descriptions and indexing as well as advanced forms of user interfaces. Agent-based technology (for both the consumer and provider) will be used to enable advanced search and retrieval functionalities. Automatic and semi-automatic AV analysis tools will be developed and integrated into a COntent Provider's Application (COPA) to enable advanced forms of AV indexing.

One of the general objectives of DICEMAN is to contribute to the emerging technologies which play a key role in the project's vision for content exchange. For this reason, project members are very active in the MPEG-7 standardisation process. In fact, the general project requirements are very close to those of MPEG-7 [4]. As such, the technologies developed within the project have been proposed in response to the MPEG-7 Call for Proposals (CfP). These include a DICEMAN Description Definition Language (DDL), audio and visual DSs and various audio Ds. This paper focuses on the DSs developed for describing still images and video sequences. The other aspects of the project are outside the scope of this paper, but the reader is referred to [5] and [6] for more details.

III Designing the DICEMAN DSs

The two DSs presented in this paper are designed based on an analogy with two well-known methods for describing written documents: a *Table of Contents* and an *Index*. A Table of Contents (TOC) is a hierarchical representation of the 1-D linear structure of a book (see Figure **¡Error! Argumento de modificador desconocido.**). It specifies important elementary components of the book (i.e. chapters, sections, sub-sections) in a hierarchical manner and assigns each a single reference (i.e. a page number) to their location in the book. Whilst the titles of the components

[1] Distributed Internet Content Exchange using MPEG-7 and Agent Negotiations

may carry semantic information, the main purpose of the TOC is to describe the structure of the book. The purpose of an Index, is to describe the contents of a book. It consists of a collection of items with references to their location in the book (see Figure **¡Error! Argumento de modificador desconocido.**). The items are selected based on their semantic value to a human reader. Clearly, a given item may appear in several parts of a book and this is handled by multiple references for the item. Very often the Index is structured hierarchically in order to allow efficient access to an item of interest.

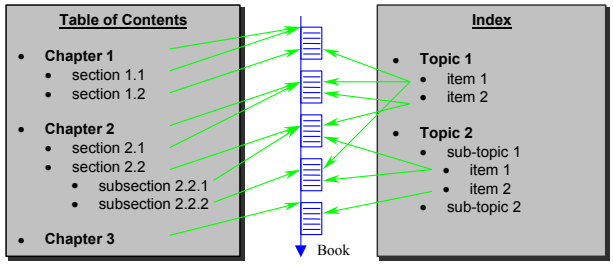


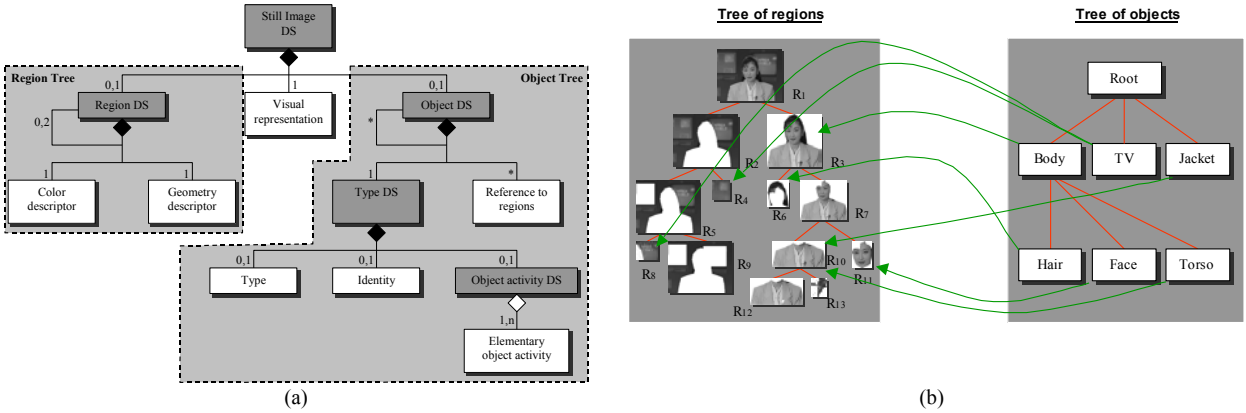
Figure **¡Error! Argumento de modificador desconocido.** - Description of a written document by a TOC and an Index

IV The DICEMAN Still Image DS

Using the analogy of the previous section, the DICEMAN Still Image DS provides a description of both the structure and the content of an image [7]. A global view of the DS is presented in Figure **¡Error! Argumento de modificador desconocido.**(a). This figure illustrates how the Still Image DS is composed of a number of sub-DSs and Ds. The numbers on each branch of the diagram represent the number of components possible in each case. In the following, the most important aspects of the overall DS are explained.

IV.1 The Region Tree (Region DS)

The goal of the Region Tree is to describe the spatial organisation of an image - it is the TOC of the image [7].A Binary Partition Tree (BTP) [9] is used to describe the inclusion relationships between sub-regions of the image. The BTP is built based on a bottom-up merging process starting with a fine region-based segmentation of the image. After merging, the BTP represents a hierarchical multiscale segmentation of the image. Large image regions are represented on higher levels of the tree, whereas fine details can be obtained from lower levels (see Figure **¡Error! Argumento de modificador desconocido.**(b)). Descriptors for signal-based features of an image region (e.g. colour, spatial and geometrical characteristics) are attached to the corresponding node in the tree. For the most part, the Region Tree can be created automatically using existing analysis tools (although user interaction is not ruled out for correction/modification).



(a) (b)

(a) Global view of the Still Image DS - grey (white) rectangles represent DSs (Ds)
(b) The Region Tree and Object Tree

Figure **¡Error! Argumento de modificador desconocido.** - The DICEMAN Still Image DS

IV.2 The Object Tree (Object DS)

The Object Tree is an arbitrary tree created by the user in order to describe the semantic content of the image in terms of objects - it is the Index of the image [7]. Each node of the tree includes a Type sub-DS which provides semantic information about the object (i.e. type, identity and activity) via references to appropriate thesauruses.

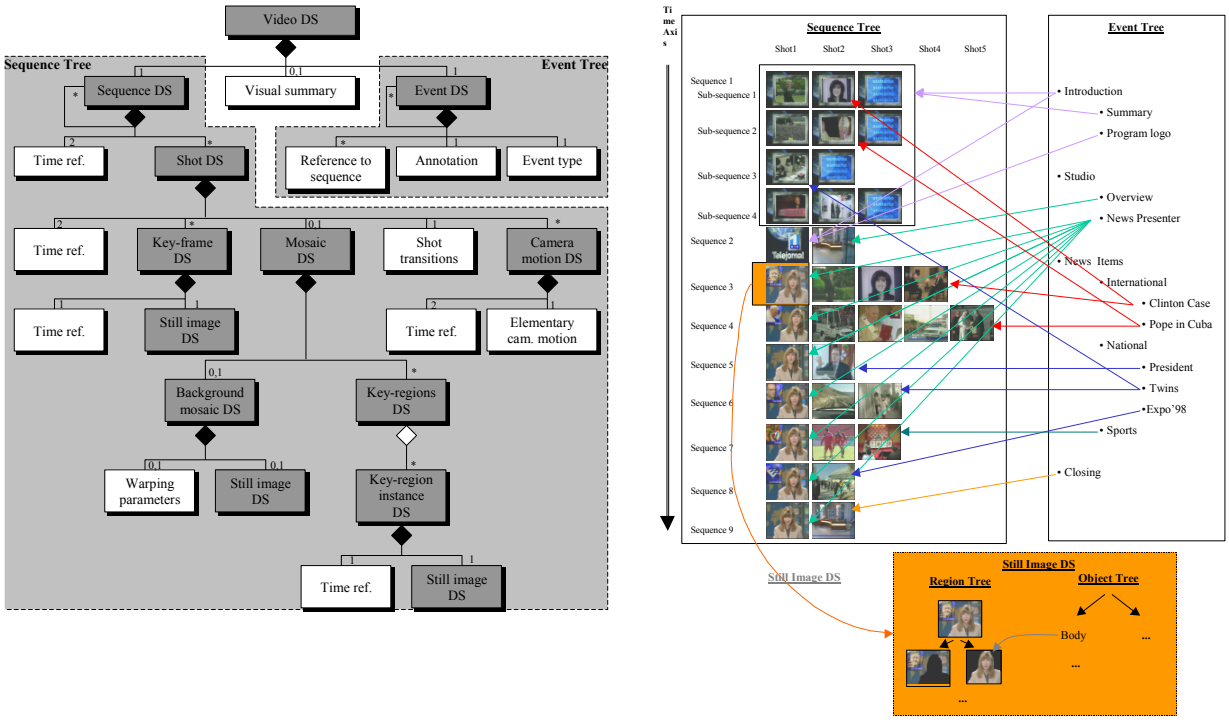
Clearly, there is a strong connection between a semantic description of an image and its signal-based properties. This is acknowledged by including a "reference to regions" descriptor at each node in the Object Tree which allows linking of semantic and signal-based descriptions (see Figure **¡Error! Argumento de modificador desconocido.**(b)). In manner similar to an Index of a written document, each node can have multiple references (e.g. in Figure **¡Error! Argumento de modificador desconocido.**(b) the TV appears in more than one image region).

IV.3 Visual Representation

The visual representation descriptor is included in the DS for browsing applications and simply corresponds to a sub-sampled version of the original image [7].

V The Video DS

The DICEMAN Video DS also provides a description of both the structure and the semantic content of a video sequence [8]. A global view of the DS is presented in Figure **¡Error! Argumento de modificador desconocido.**(a). As in the Still Image DS, the Video DS is composed of a number of sub-DSs and Ds. In the following, the most important aspects of the overall DS are described.



(a) (b)

(a) Global view of the Video DS - grey (white) rectangles represent DSs (Ds)

(b) An example of the Sequence and Event Trees

Figure **¡Error! Argumento de modificador desconocido.** - The DICEMAN Video DS

V.1 The Sequence Tree (Sequence DS)

The Sequence Tree describes how a video clip can be sub-divided into temporally connected components referred to as "sequences" - it is the TOC of the video clip (see Figure **¡Error! Argumento de modificador desconocido.**(b) and [8]). The leaves of the tree are assumed to be shots (i.e. "sequences" which do not involve any editing effects). The descriptor associated to non-leaf nodes of the tree is simply a time reference descriptor indicating the beginning and the end of the corresponding sequence. The description of shots (i.e. the Shot DS) includes a time reference descriptor, a transition descriptor (e.g. cut, fade, wipe), a camera activity DS which describes the type of camera work used (e.g. tilt, pan, zoom) and several visual representation sub-DSs. The keyframe DS specifies a number of images in the shot which are representative of the action taking place [10]. The mosaic DS includes the background mosaic DS which describes how every image in the shot can be warped to a common reference in order to construct a representative background mosaic image [11]. Conversely, the key-region DS describes the objects moving across this background mosaic. Each of these visualisation DSs consists of a time reference descriptor and the Still Image DS as described in section IV. For the most part, the Sequence Tree can be created automatically using existing analysis tools.

V.2 The Event Tree (Event DS)

The Event Tree is an arbitrary tree created by the user which hierarchically defines the main events taking place in the video clip - it is the Index of the video clip (see Figure **¡Error! Argumento de modificador desconocido.**(b) and [8]). It includes descriptors related to the semantic of what is happening in the video. Many events can be pre-defined in a thesaurus and thus the required descriptors can simply consist of an index defining the event type. Alternatively, annotation could be used. Finally, in order to relate events with temporal video segments, descriptors called "reference to sequence" are assigned to each event.

V.3 Visual Summary

The visual summary descriptor is included in the DS for browsing applications and simply corresponds to a temporally and spatially sub-sampled version of the original video clip [8].

VI Conclusion

The presented DSs received a very favourable evaluation in the MPEG-7 CfP evaluation process [12]. As such, they have been selected as the basis for further development within the standardisation process. The objective is to use these DSs as a starting basis and to integrate other DSs, which also received favourable evaluations, but which address weaknesses[2] of the two DSs (or functionalities[3] not addressed). This enhanced and more complete visual DS will then be integrated into the development of a generic MPEG-7 audio-visual DS [13]. In parallel, the first version of COPA has recently been developed within DICEMAN and features automatic analysis tools for shot cut detection, keyframe extraction (based on [10]), region tree generation (based on [9]), camera motion parameter estimation, and warping parameter estimation, which produce descriptions conforming to the presented DSs. In the next phase of COPA development, additional tools will be added in order to produce complete descriptions (e.g. manual annotation, object segmentation and tracking). Modifications will then be made to produce a final version of COPA which produces descriptions conforming to a subset the generic DS developed by MPEG-7.

Acknowledgement

The achievements described in this document were carried out within the ACTS project DICEMAN. The work was part-funded by the European Commission. The views expressed are those of the authors and should not be taken to represent, in any way, the views of the European Commission or its services

References:

[1] MPEG Requirements Group, "MPEG-7 Context, Objectives and Technical Roadmap", Doc. ISO/IEC JTC1/SC29/WG11 N2729, Seoul meeting, March 1999.
[2] MPEG Requirements Group, "MPEG-7 Applications Document", Doc. ISO/IEC JTC1/SC29/WG11 N2728, Seoul meeting, March 1999.
[3] MPEG Requirements Group, "MPEG-7 Proposal Package Description ", Doc. ISO/IEC JTC1/SC29/WG11 N2464, Atlantic City meeting, October 1998.
[4] MPEG Requirements Group, "MPEG-7 Requirements Document", Doc. ISO/IEC JTC1/SC29/WG11 N2727, Seoul meeting, March 1999.
[5] DICEMAN Consortium, *The DICEMAN WWW Site*, http://www.teltec.dcu.ie/diceman
[6] DICEMAN Consortium, "DICEMAN Initial Requirements", Public Deliverable, available on the Internet: http://www.teltec.dcu.ie/diceman
[7] DICEMAN / France Telecom, "Proposal for MPEG-7 evaluation: Still image DS", Technical Report ISO / IEC JTC1 / SC29 / WG11 / P186, Lancaster evaluation meeting, February 1999
[8] DICEMAN Consortium, "Proposal for MPEG-7 evaluation: Video DS", Technical Report ISO / IEC JTC1 / SC29 / WG11 / P185, Lancaster evaluation meeting, February 1999
[9] P. Salembier, L. Garrido, "Binary partition tree as an efficient representation for filtering, segmentation and information retrieval", IEEE ICIP'98, Chicago (IL), USA, Oct. 4-7, 98.
[10] A. Hanjalic, R.L. Lagendijk, and J. Biemond, "A New Key-Frame Allocation Method for Representing Stored Video-Streams", First International Workshop on Image Databases and Multi Media Search, Amsterdam, The Netherlands, 1996.
[11] H. Sawhney, S. Ayer. "Compact representations of videos through dominant and multiple motion estimation", IEEE Trans. on Pattern Analysis and Machine Intelligence, 18:814-830, August 1996
[12] MPEG Requirements Group, "MPEG-7 Evaluation results", Doc. ISO/IEC JTC1/SC29/WG11 N2730, Seoul meeting, March 1999
[13] MPEG Convenor, "AHG on MPEG-7 description schemes", Doc. ISO/IEC JTC1/SC29/WG11 N2735, Seoul meeting, March 1999.

[2] E.g. richer semantic descriptions allowed the definition of entities and relationships between entities etc.
[3] E.g. allowing descriptions of synthetic and/or 3-D visual content (MPEG-4 BIFS, etc.).