Using the MPEG-7 Audio-Visual Description Profile for 3D Video Content Description

Nicholas Vretos, Nikos Nikolaidis, Ioannis Pitas

Computer Science Department, Aristotle University of Thessaloniki, Thessaloniki, Greece

E-mail: {vretos, nikolaid, pitas}@aiia.csd.auth.gr

Abstract: **In this paper we propose a way of using the Audio-Visual Description Profile (AVDP) of the MPEG-7 standard for stereo video content. Our aim is to provide means of using AVDP in such a way that 3D video content can be correctly and consistently described. Since, AVDP semantics do not include ways for dealing with 3D video content, and thus, a new semantic framework within AVDP is proposed. Finally, we show some examples of xml files that describe stereo video content.**

Keywords: Stereo video, MPEG-7, AVDP.

1 INTRODUCTION

Automatic analysis of videos consists of algorithms for shot boundaries detection, face detection/tracking/recognition, facial expression recognition and others. These algorithms are used in applications such as fast indexing in databases, implementation of better editing tools, as well as better program schedulers in real-time applications such as the TV context. It is very easy to conclude that video data increase exponentially with time and researchers are focusing on finding better ways in classification and retrieval applications for video databases. Moreover, the way of constructing videos has also changed in the last years. The potential of digital videos gives producers better editing tools for a film production. Finally, automatic schedulers for TV programming are essential in the broadcasting business.

The new trend in multimedia is the use of 3D representation. Most of the recent film productions have their 3D versions. Stereo video is an approach of 3D video film making, based on the human eye system [1]. It consists of two different cameras put together side-by-side (most of the time), and therefore, by means of special glasses and viewing software, the viewer is able to perceive a 3D motion picture in front her/his eyes. Analysis of stereoscopic video have the advantage of additional information to improve results of the before mentioned algorithms, and also, derive annotation for 3D specific content such as 3D position of foreground objects, viewing quality of 3D content and others.

For a better manipulation of all the above, MPEG-7 standardizes a set of Descriptors (Ds), Description Schemes (DSs), a description definition language (DDL) and a description encoding [2]-[6]. A considerable amount of research effort, have been invested over the last years to improve MPEG-7 performance in terms of semantic content description [7]-[10]. Nevertheless, 3D content description has not yet been investigated in the MPEG-7 context. Although some description and description schemes have been proposed to model 3D information, they are only explicit descriptors for geometrical information and not for 3D video content.

The AudioVisual Description Profile (AVDP) is very recently adopted as a new profile of the MPEG-7 standard. This profile consists of a subset of the original MPEG-7 standard, and aims in describing the results of most of the known media analysis tasks (e.g. shot detection, face detection/tracking, and others), in a normative manner.

Our aim is to show that AVDP can be used, with new semantics, to describe also 3D media analysis tasks results. An approach of using the descriptors and description schemes of the AVDP is proposed, for most of the known media analysis tasks extended to the 3D context.

The paper is organized as follows: Section 2, an overview of the AVDP will be presented and the way 3D information can be incorporated. In Section 3, we show the details of the 3D AVDP semantics for several known media analysis algorithms. In Section 4, specific 3D issues that need to be integrated in the AVDP are treated. Finally, in Section 5, conclusions and future work is discussed.

2 THE AUDIOVISUAL DESCRIPTION PROFILE (AVDP)

The Audio Visual Description Profile (AVDP) of MPEG-7 provides a standard way to store high and low level information, which is extracted from the content analysis of video. AVDP was designed to benefit both broadcasters and industry in order to create a normative layer between research and end users (i.e. TV broadcasters and media analysis industry). In this paper, we propose to store semantic analysis results to an XML file. This XML file must be compatible to the specifications described in the XML Description Schema (XSD) of the AVDP. We have selected a subset of the description tools available in the AVDP, which cater to our needs, and we have defined a description procedure, using these description tools, in order to store information describing 3D content.

Corresponding author: Nicholas Vretos, Computer Science Department, Aristotle University of Thessaloniki, vretos@aiia.csd.auth.gr

In the following a 3D video segment can mean one of the following:

- a stereoscopic video consisting of two channels (left and right)
- a stereoscopic video consisting of two channels (left and right) and two or four extra channels containing corresponding disparity information (horizontal and vertical)
- a video consisting of a color channel and a depth information channel.

The following list contains most of the tools and the description schemes (DSs) that were used to our end.

- TemporalDecomposition (TD). This description scheme (DS) acts as a tool that performs temporal decomposition of the video in multiple temporal segments, such as Video Segments (VS) or Audio-Visual Segments (AVS).
- MediaSourceDecomposition (MSD): used in the AVDP context to decompose an audiovisual segment (or an entire audiovisual content) into the audio and video channels that it contains.
- VideoSegment (VS): provides a way to describe a video segment of the visual content. The starting time point of the VS and its time span defines each segment.
- SpatioTemporalDecomposition (STD): enables a video segment to be decomposed into parts defined spatiotemporally namely MovingRegions (e.g., in order to store information related to moving objects).
- MovingRegion (MR): used to describe, for example, a moving object by storing the spatiotemporal behavior of the object (spatial coordinates of the bounding box of an object and their change with respect to time).
- SpatialDecomposition (SD): used for the spatial decomposition of a frame. The result is one or more regions of interest (StillRegion) within a frame that may depict an object, face etc.
- StillRegion (SR): used to describe a still object, by defining its spatial span within a frame. A StillRegion can also denote an entire frame.
- StructuredAnnotation: used to annotate concepts, events, human actions etc.
- FreeTextAnnotation. Annotate a video segment with free text.

In order to stay consistent with the AVDP profile, we use different content entities for each video and depth channel. A content entity, as its name indicate, is a container type able to store all lower level descriptors of the AVDP. It is used a the root of the description for a specifc channel. By these means, we treat each channel as a different video. Figure 1 below illustrates this fact.

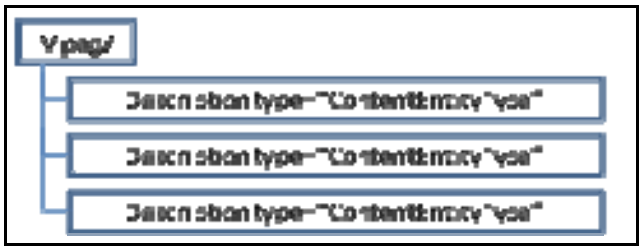
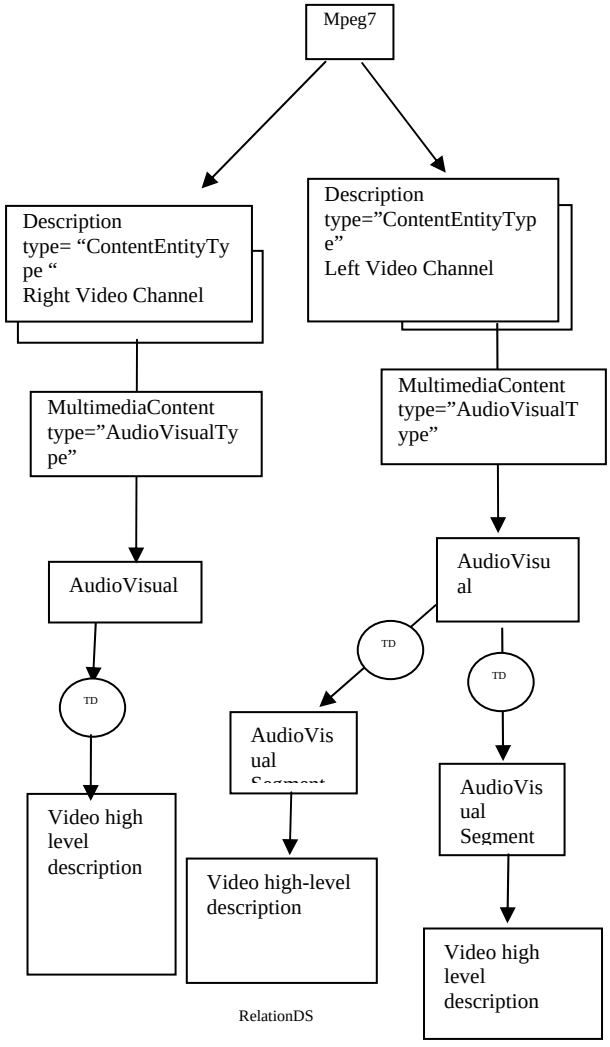


Figure 1: Each contentEntityType is a different video channel or depth channel.

Therefore, we create for each ContentEntityType, the analysis tree based on the AVDP as show in Figure 2.



Relation between object appearing in left and right channels.

Figure 2: AVDP tree for 3D Video description

3 AVDP FOR 3D VIDEO ANALYSIS TASKS

In this section we shall describe 10 different 3D content analysis algorithms and the way their results can be described using the AVDP. The analysis tasks that we deal with, are major research areas in the video and 3D video processing area. The algorithms are:

- Scene boundaries detection
- Shot boundaries detection
- Key Frames and Key Video Segment Extraction
- Object Detection
- Object Tracking
- Human Activity Recognition
- Face Clustering
- Object Clustering
- Facial Expression Recognition

3.1 Scene/Shot Boundaries Detection

A scene/shot boundaries detection algorithm is able to detect boundaries of scenes/shots in multimedia content. The aim is a temporal decomposition of a multimedia content into different scenes/shots. For the scene/shot boundaries detection algorithm a temporal decomposition of an AudioVisualSegment is generated. The description of such an output is simple in terms of AVDP.

Instead, a scene or shot boundaries detection algorithm, may store its results in an AudioVisual Temporal Decomposition (TD). Each resulting AVS will include some temporal information and nothing else. The schematic representation of such a descriptor is shown in Figure 3.

Moreover, with exactly the same approach we can handle the cases of shot transitions, which exceed the simple cut case. Such transitions are the fade-in, fade-out, dissolve and others, which contains more than one frame. In these cases we use the same description scheme and code the transition as a shot (i.e. a Video Segment).

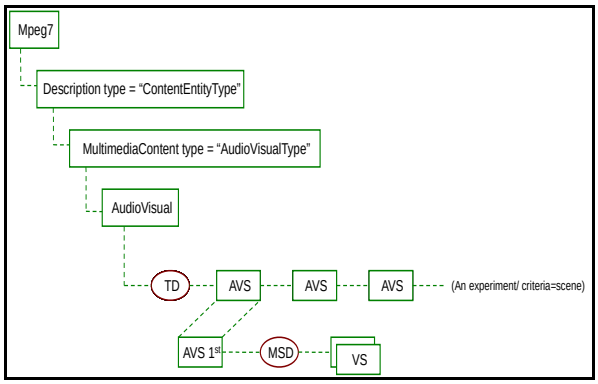


Figure 3: Scene/shot schematic representation

3.2 Key Frames and Key Video Segment Extraction

Key frames extraction refers to the multimedia analysis task, where some characteristic frames are extracted from a video segment (in most cases from a shot). Key video segments extraction is the analysis task where characteristic video segments, namely visual summaries, are extracted from a shot. Key frames and segments can be used for fast browsing and condensed representation of query results in a 3D video asset management environment. The key frames and key video segments extraction algorithm generates a list of video segments of duration of one frame or more frames, respectively. The description of such an output is simple in terms of AVDP. The schematic representation of such a descriptor is shown in Figure 4.

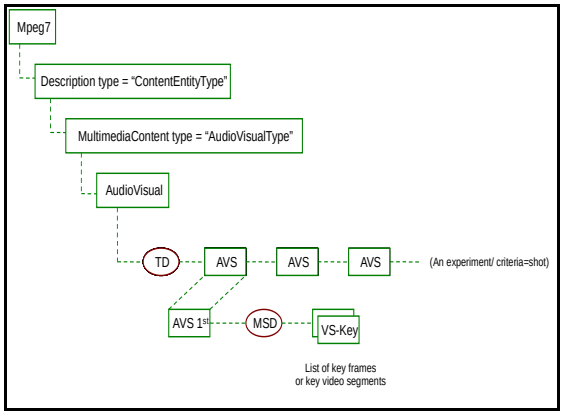


Figure 4: Key Frame and Key Video Segment schematic representation

3.3 Object Detection/Tracking

Object detection is the process of finding a predefined object (e.g. a face, a car, a ball etc) in a 3D video. Usually, object detection is performed in a per-frame basis, although an extension for object detection on a video segment using a frame-by-frame approach is straightforward. Since the object detector usually detects a specific object or a specific category of objects, this information can be used to semantically annotate the detected object with its type. For instance, a face detector detects faces and thus we know that all objects detected by such a detector are faces and we can store this information within the object. On the other hand, object tracking is the process of finding the trajectory of a predefined object (e.g. a face) in a sequence of frames. Usually, object tracking is performed in a video segment in a frame-by-frame basis where the spatial position of the tracked object (usually in the form of a bounding box) is calculated for each frame. The results of such process are the object trajectory.

In the case of an object detection module the AVDP profile provides us with a StillRegionDS which can be used to store the location of the detected object(s) (usually in terms of a bounding box) as well as other relevant information for this region. This is presented in Figure 5.

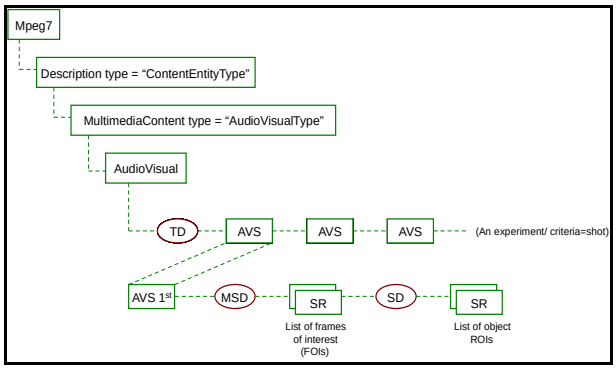


Figure 5: Object Detection schematic representation

In the AVS-1st level of the above description, we use a MediaSourceDecompositionDS (MSD) to decompose the VS into a list of StillRegionType elements where each of them represents an entire frame of the video segment. This list can be considered as the frames of Interest (FOIs). Subsequently, we decompose each frame into further StillRegionsDS through a Spatial Decomposition DS, each StillRegion representing a detected object.

To store the results of object tracking algorithms, the AVDP profile provides us with a MovingRegionDS (MR), which can be used to store information regarding a spatial region (e.g. a bounding box) that moves over time.

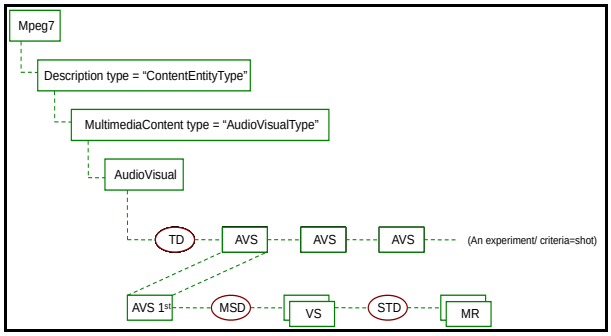


Figure 6: Object tracking schematic representation

It has to be noted that, as in the case of object detection, the tracker may give a first identification of the tracked object through the criteria of the spatiotemporal decomposition description scheme as well as its structural units. This is possible only if an object detection algorithm is used to initialize the tracking algorithm.

3.4 Human Activity Recognition

Human activity recognition aims at recognizing specific, predefined human activities in a video segment (i.e. running, walking, sit down etc). The video segment may be an entire shot or a video frame within a shot.

In the case of human activity recognition an annotation of a StillRegionDS or a MovingRegionDS is generated. Since both types derive from the SegmentDS type and since the AVDP profile includes the StructuredAnnotation of the SegmentDS, we can use it to provide human activity recognition description. Thus in both cases we use the WhatAction tag of the StructuredAnnotationDS to tag the recognized activity. Since we might have more than one actions taking place in a MovingRegionDS, we have to provide a way to be able to describe such different actions within the same MovingRegionDS. To do so, we can further decompose the initial MovingRegionDS into its semantically meaningful entities, i.e. moving regions, each representing a single activity. In more detail, we use the MovingRegionTemporalDecompositionType and thus create MovingRegionDSs (without the intialRegion tag) in order to characterize the activities occurring in different segments of the initial MovingRegionDS. It has to be noted, that we can allow overlapping MovingRegionDSs within the MovingRegionTemporalDecompositionType for different activities that can take place in the same time. This is useful to describe both the activity and the facial expression of a person, since (see Section 3.6) essentially we consider affect display as an activity. For instance, if we are able to characterize frames 1 to 10 of a video segment with action "Walk", while in the same time we can characterize frames 5 to 10 with an affect action of "Happiness" then we create two different MovingRegionDSs within the MovingRegionTemporalDecompositionType of the initial MovingRegionDS (containing the tracking information). These two MovingRegionDSs will overlap for frames 5 to 10. Moreover, if we have only one characterization for the entire initial MovingRegionDS (i.e. the same action is performed in the entire moving region), such decomposition can be discarded and we can directly characterize the initial MovingRegionDS. Both descriptions are valid and consistent with the AVDP profile.

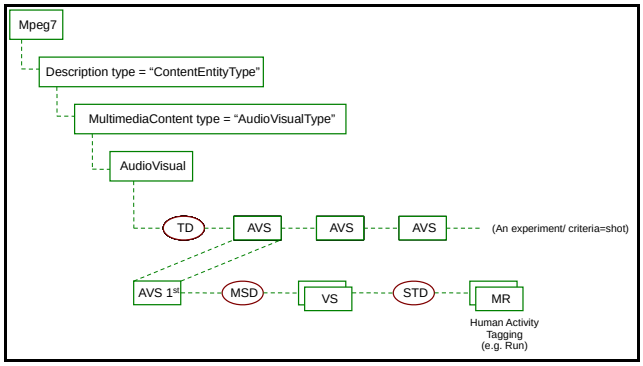


Figure 7: Human Activity Recognition schematization

Moreover, in cases where extra information about the human activity may be extracted (e.g. by geometric reasoning over the trajectory of the person), we can use the How tag of the structured annotation to describe this information as well. Such information may be for instance the direction of the activity as "Left" or "Right" (e.g. for walking).

3.5 Face/Object Clustering

Face/Object Clustering is defined as the analysis task, which partitions different facial or object images into clusters based on the actor or object they represent. In the case of 3D video, video frame regions representing faces or objects can be clustered into clusters of actors/objects. For storing the results of face/object clustering, we simply update appropriately the Who/WhatObject tag

respectively in the structured annotation of each involved segment type (i.e. MovingRegionDS or StillRegionDS). The values that are associated with Who/WhatObject tags are mainly values such as “Face”, “Car”, “Actor_1” etc.

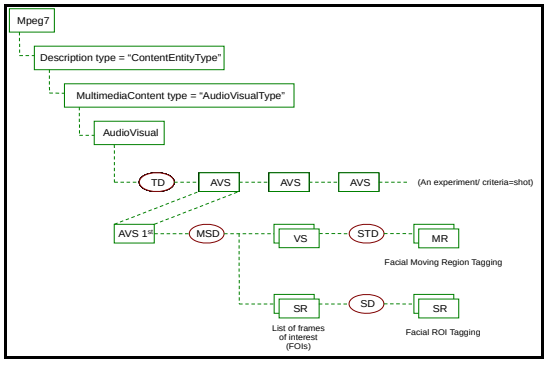


Figure 8: Face/Object Clustering schematic representation

3.6 Facial Expression Recognition

Facial expression recognition is used in order to recognize predefined facial expressions such as happiness, anger, fear etc. In the case of Facial Expression Recognition an annotation of a StillRegionDS that defines the image area (bounding box) where a face is depicted is needed. A WhatAction tag with value “affect” is used within the StructuredAnnotation whereas the How tag is used to label the recognized expression.

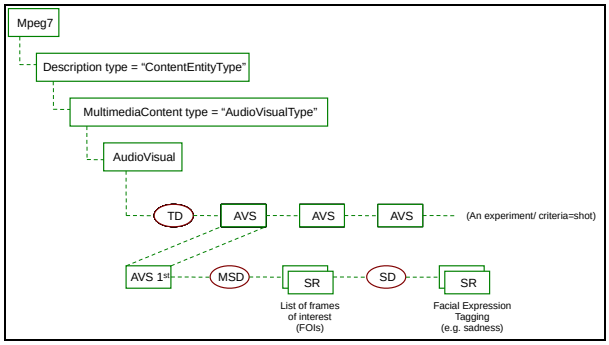


Figure 9: Facial Expression schematic representation

4 DEALING WITH 3D VIDEO SPECIFIC ISSUES IN AVDP

4.1 Correspondences and Discrepancies Between Still or Moving Regions in Two or More Channels

When Object/person detection or tracking algorithms, are applied on stereo data (left and right video channels) or stereo data and their respective disparity channels may result in the derivation of correspondences between still regions (e.g. bounding boxes) or moving regions (e.g. bounding boxes in successive frames that correspond to the same object or person) in two channels, e.g. between the right and left video channel. These correspondences denote that the two still regions or moving regions depict the same physical entity (object or person) in the two channels. For example, an object detection algorithm may use the two video channels and detect the depictions of the same physical object in these channels. Thus, this relation/correspondence should be encoded in the XML file. To do so, we use the RelationDS from the AVDP profile. This description scheme allows us to connect two different SegmentDS types (SegmentDS is the abstract parent type for VideoSegmentDS, MovingRegionDS, StillRegionDS as well as many other decomposition types). Typically, the RelationDS type will have a strength component showing the strength of the specific relation, as well as an annotation such as “dialogue”, “handsake” and other to show the nature of the relation.

Furthermore, discrepancies (mismatches) between the content of corresponding (moving or still) regions in the various channels can also be described. Colorimetric mismatch is such a discrepancy. The RelationDS is used again, in order to designate the type of the discrepancies between two such regions. To this end, a RelationDS is inserted in the MovingRegionDS and StillRegionDS instantiation, where the type of the RelationDS designates the type of discrepancy. Note that, there can be many types of discrepancies for a specific region pairs (moving or still).

4.2 Storage of analysis results to video channels, consistency/inconsistency between analysis results derived from different channels

The availability of 2 (left+right or video+depth) or 4 (left + right + left horizontal disparity + right horizontal disparity) or 6 (left + right + left horizontal and vertical disparity + right horizontal and vertical disparity) video channels poses the question of how to analyse them and where to store the derived information. The rule that we have adopted is the following: The derived information will be stored in one, more, or all channels, depending on the algorithm. For example if a tracking algorithm is applied on the disparity channels only, the derived moving regions will be stored in these channels and the video (color) channels may contain no moving regions. Alternatively, one may choose to “copy” the moving regions derived from the disparity channels to the video (color) channels. In the particular case of information related to depth (e.g. extent/position in the z/depth axis of an object), this will be stored only in the disparity/depth channels, since it can be derived only if such channels are available.

Furthermore, in some cases, we may have conflicting results from an algorithm when applied into different channels of the same video. Such cases might arise for example in human activity recognition. If such an algorithm is applied on each channel (left/right) separately, it may derive (due to errors) different activity characterizations (e.g. walking vs. running) of the same segment in the two channels. In this case there are two options, both of which can be adopted:

a) The analysis algorithm itself combines the results stored in the two channels (in a subsequent fusion step) and tags both channels with the same consistent tag using the appropriate AVDP descriptors in the XML file.

b) There is no combination of results in the two channels and each of them is tagged (inconsistency) with a different tag in the XML file.

4.3 Geometrical Reasoning

By geometrical reasoning we define the ability to annotate 3D video content with information related to the geometric properties or relations of objects/ persons. Such properties may refer to the geometrical position of an actor in the 3D world, the geometrical proportions (size) of an actor/object, the spatial relation between two objects (near or far), the speed and direction of movement of an object or person etc. In order to derive such annotations the location/motion information derived from detection and tracking algorithms should be processed.

Tagging objects or persons with semantic concepts, like big/fast etc., requires in most cases a set of rules and thresholds. For example, in order to characterize an object as being centrally located in screen space, a rule (e.g. "the distance of the center of mass of the object from the screen center should be smaller than threshold T") and a threshold value (e.g. T=0.2) are required. In a more general case, tagging might require an algorithm along with its parameters.

It has to be noted that although depth information is very helpful for 3D semantic content analysis, the extraction of specific geometry related information can be performed even when absolute depth information is not available (hence, only relative depth is known), but also when there is total lack of depth information. Absolute depth information can be extracted if the stereo camera parameters are known. In case camera parameters are unknown, relative depth and, hence, geometry related information can be inferred based on the simple fact that the disparity is inversely proportional to depth. In this way, geometry related information on static objects, can be calculated using the disparity values of objects, as well as the overall disparity range. Similarly, for moving objects, disparity values can be used to obtain relative 3D trajectory of the object (e.g. moving towards the camera, term 2.30). Depth related tags will be stored in the disparity/depth channels only, provided that such channels are available.

The AVDP descriptors where such geometric descriptions will be stored are StillRegionDS, MovingRegionDS and VideoSegmentDS for still and moving objects/actors. StillRegionDS can refer either to an entire video frame or to a spatial region within a frame. We identify the following 4 description cases: a) one still object (alternatively: object properties in a single time instance), b) many objects ROIs, c) one moving object (alternatively: an object behavior during a time interval) d) many moving objects.

5 CONCLUSIONS

In this paper we have presented a new way of using the AVDP profile of the MPEG-7 standard for 3D video content analysis. We have detailed how several known algorithms can be described within such a context and how specific 3D metadata can be incorporated in the schema. The derived description can be used for storing the analysis results in a multimedia database (e.g., a 3DTV content database) or a media asset management system. Such a database can then be queried with high level queries such as "Return the video where such an actor is near the screen" in a 3D context. Finally, note that the extension of the proposed XML description method in order to support video coming from a multi-view setup (i.e., multiple cameras) is straightforward.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 287674 (3DTVS). The publication reflects only the authors' views. The EU is not liable for any use that may be made of the information contained therei.

References

[1] Smolic, A.; Kauff, P.; Knorr, S.; Hornung, A.; Kunter, M.; Müller, M.; Lang, M.; , "Three-Dimensional Video Postproduction and Processing," *Proceedings of the IEEE* , vol.99, no.4, pp.607-625, April 2011.

[2] I.S.O, "Information technology – multimedia content description interface - part 1: Systems," , no. ISO/IEC JTC 1/SC 29 N 4153, 2001.

[3] I.S.O, "Information technology – multimedia content description interface - part 2: Description definition language," , no. ISO/IEC JTC 1/SC 29 N 4155, 2001.

[4] I.S.O, "Information technology – multimedia content description interface - part 3: Visual," , no. ISO/IEC JTC 1/SC 29 N 4157, 2001.

[5] I.S.O, "Information technology – multimedia content description interface - part 4: Audio," , no. ISO/IEC JTC 1/SC 29 N 4159, 2001.

[6] I.S.O, "Information technology – multimedia content description interface - part 5: Multimedia description schemes," , no. ISO/IEC JTC 1/SC 29 N 4161, 2001.

[7] John P. Eakins, "Retrieval of still images by content," pp. 111–138, 2001.

[8] A. Vakali, M.S. Hacid, and A. Elmagarmid, "Mpeg-7 based description schemes for multi-level video content classification," IVC , vol. 22, no. 5, pp. 367–378, May 2004.

[9] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," Computer Vision and Image Understanding , vol. 73, no. 3, pp. 428–440, 1999.

[10] J.J. Wang and S. Singh, "Video analysis of human dynamics - a survey," Real-Time Imaging , vol. 9, no. 5, pp. 321–346, October 2003.