Human-centered 2D/3D Video Content Analysis and Description

K. Papachristou,[1] N. Nikolaidis,[1,*] I. Pitas,[1,†] A. Linnemann,[2] M. Liu,[2] and S. Gerke[2]
[1]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2]Image Processing Department, Fraunhofer Institute HHI, Berlin, Germany
{*nikolaid,†pitas}@aiia.csd.auth.gr

***Abstract*—In this paper, we propose a way of using the AudioVisual Description Profile (AVDP) of the MPEG-7 standard for stereo video and multichannel audio content description. Our aim is to provide means of using AVDP in such a way, that 3D video and audio content can be correctly and consistently described. Since AVDP semantics do not include ways for dealing with 3D audiovisual content, a new semantic framework within AVDP is proposed and examples of using AVDP to describe the results of analysis algorithms on stereo video and multichannel audio content are presented.**

***Index Terms*—AudioVisual Description Profile (AVDP), MPEG-7, stereo video, semantic content description**

I. Introduction

Automatic analysis of video and audio includes various tasks e.g., shot/scene boundary detection, person detection/tracking/recognition, facial expression recognition, music/speech segmentation, speaker diarization and music genre/ mood characterization. The relevant algorithms are used in various applications, such as semantic description of video content for archival, indexing and retrieval, implementation of better audiovisual editing tools, intelligent content manipulation, etc. Recorded, broadcasted and webcasted data increase exponentially with time and research is focused on finding ways to analyse, describe and annotate video content in an efficient and automatic manner.

A rather new trend in multimedia is the use of stereoscopic video. Many of the recent film productions have their 3D versions [1]. Analysis of stereoscopic video has the advantage of benefiting from the additional available information, namely depth/disparity, which can boost the performance of analysis algorithms, such as the ones mentioned above. In addition, analysis of stereoscopic video can derive information that cannot be inferred from single-view video, such as 3D object position. Finally, the particularities of 3D video call for analysis algorithms that can characterize 3D video quality or its conformance to the rules of 3D cinematography [2].

For handling audiovisual content annotation and description, MPEG-7 standardizes a set of Descriptors (Ds), Description Schemes (DSs), a description definition language (DDL) and a description encoding [3]. A considerable amount of effort has been invested over the last years to improve MPEG-7 ability to deal with semantic content description. Nevertheless, 3D audiovisual 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) has been recently adopted as a new profile of the MPEG-7 standard [4]. This profile consists of a subset of the original MPEG-7 standard and aims at describing the results of most of the known audiovisual analysis tasks (e.g., shot detection, face detection/tracking), in a normative manner. AVDP was designed to benefit both broadcasters and the digital media industry in order to create a normative layer between media content production and consumption. In this paper, we propose to store 3D video and multichannel audio content (e.g., 3DTV content) semantic analysis results to an XML file compatible to the specifications of the AVDP. Our aim is to show that AVDP can be used, by properly utilizing its descriptors and description schemes, for describing and storing the results of 3D audiovisual content analysis.

II. The Use of the AudioVisual Description Profile for 3D Content Description

The AudioVisual Description Profile (AVDP) provides a normative way to store high and low level information, extracted from the analysis of video and/or audio content. We have selected a subset of the decomposition types (namely TemporalDecomposition (TD), MediaSourceDecomposition (MSD), SpatioTemporalDecomposition (STD), SpatialDecomposition (SD)) available in the AVDP, in order to create various segment types (AudioVisualSegment (AVS), VideoSegment (VS), AudioSegment (AS), StillRegion (SR), MovingRegion (MR)) for representing the output of various analysis algorithms. In our approach, a content entity is used as the root of the description for a specific channel. Such a channel can be a color or disparity channel (left/right) of a stereo video segment or the audio content description, which unlike video is stored in a single content entity despite the fact that multichannel audio is present. All the relevant information of a channel obtained by various media analysis algorithms is stored within the content entity representing the corresponding channel by employing the above mentioned decomposition and segment types. Information that does not refer to a specific video channel (e.g., left or right), but to the 3D video as a whole, is stored to a reference channel, namely left channel. In the following sections, we will describe a number of 3D video and audio content analysis algorithms and the way their results can be stored using AVDP. Figure 1 illustrates the proposed AVDP usage for 3D content description.

III. Description of 3D Video Analysis Results

A. *Scene/Shot Boundary Description*

A scene/shot boundary detection algorithm can detect boundaries of scenes/shots in 2D or stereo video content, resulting in a temporal decomposition of a multimedia content into different scenes/shots [5], [6]. In terms of AVDP, a

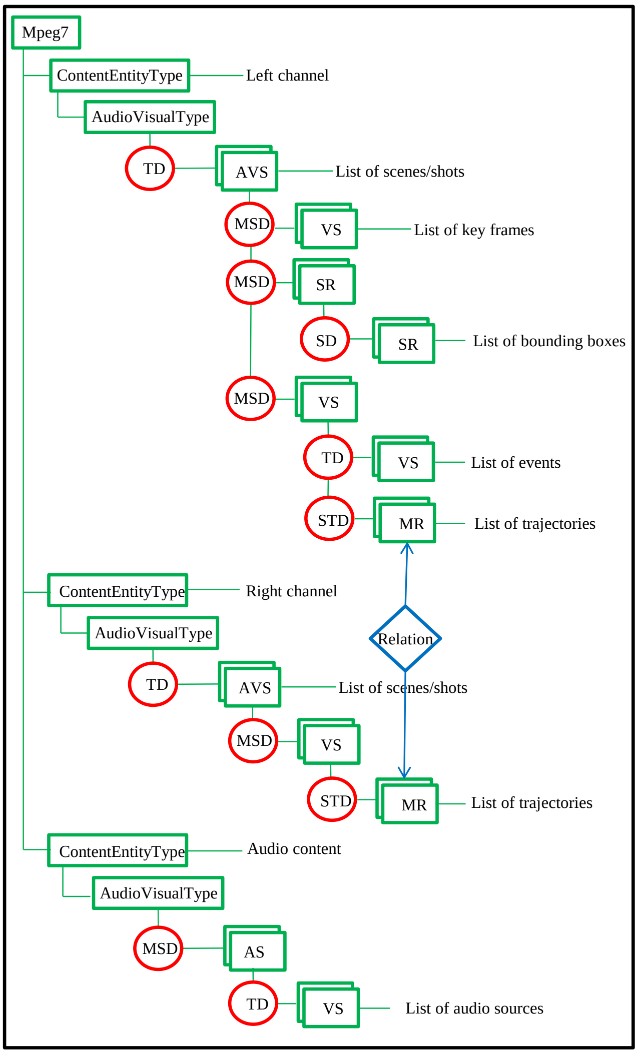

Fig. 1. Schematization of AVDP usage for 3D content description.

TemporalDecomposition of an AudioVisualSegment describing the entire channel of a 3D video content is generated. Each resulting AudioVisualSegment represents a scene/shot. Moreover, with the same approach we can handle the case of shot transitions, such as fade-in, fade-out, dissolve etc. In all the cases, the type of the segment (e.g., scene, shot, transition of a certain type) is stored using the StructuralUnit element within the AudioVisualSegment. Additionally, semantic characterizations of scenes/shots, such as comfortable for 3D viewing, wide, close-up or popup, is provided using the How element of the StructuralAnnotation type. In the following partial XML example, a shot with its time and shot type information (long shot) is described:

```
<AudioVisualSegment id="Shot_25">
 <StructuralUnit href="StructuralUnitCS#shot"/>
 <TextAnnotation>
  <StructuredAnnotation>
   <How href="SceneShotCharacterizationCS#long"/>
  </StructuredAnnotation>
 </TextAnnotation>
 <MediaTime>
  <MediaRelIncrTimePoint>25</MediaRelIncrTimePoint>
  <MediaIncrDuration>142</MediaIncrDuration>
 </MediaTime>
</AudioVisualSegment>
```

B. *Key Frame and Key Video Segment Description*

Key frame or key video segment extraction algorithms produce characteristic video frames/segments summarizing a video segment [7]. Both color and depth information can be used for 3D video summarization. Key video frames and segments provide a video summary that can be used e.g., for fast browsing of query results in a 3D video asset management system. To describe such a summary, the MediaSourceDecomposition type is used within each respective shot, in order to generate a list of VideoSegment types representing each a key frame/video segment with the corresponding frame duration. Information regarding the respective multimedia data (e.g., image files for the key frames) can be stored using the MediaLocator type. A key frame example is given below:

```
<MediaSourceDecomposition
    criteria="DecompositionCS#key_segments">
 <VideoSegment id="KeyFrame_11">
  <MediaLocator>
   <MediaUri>/kf_34.jpg</MediaUri>
  </MediaLocator>
  <StructuralUnit href="StructuralUnitCS#key_frame"/>
   <MediaTime>
    <MediaRelIncrTimePoint>34</MediaRelIncrTimePoint>
    <MediaIncrDuration>1</MediaIncrDuration>
   </MediaTime>
  </VideoSegment>
</MediaSourceDecomposition>
```

C. *Event Detection*

An event is a semantically important concept having a certain duration in the audiovisual stream, such as a dialogue among people or a car accident. For describing events occuring within a shot, we use the TemporalDecomposition type of the VideoSegment, representing the visual information of the shot. The criteria attribute is set to "events", to create a series of VideoSegments, each corresponding to a single event. The specific type of an event is stored within the StructuralUnit element of VideoSegment. An example of a dialogue occuring in the time interval [10, 50] is shown below:

```
<TemporalDecomposition
    criteria="DecompositionCS#events">
 <VideoSegment id="Event_13">
  <StructuralUnit href="StructuralUnitCS#event.dialogue"/>
  <MediaTime>
   <MediaRelIncrTimePoint>10</MediaRelIncrTimePoint>
   <MediaIncrDuration>41</MediaIncrDuration>
  </MediaTime>
 </VideoSegment>
</TemporalDecomposition>
```

D. *Semantic 3D Video Quality Descriptions*

3D video quality is highly related to the depth perception and the visual comfort while watching such a video [2], [8]. Semantic qualitative descriptions of 3D quality are more intuitive than quantitative descriptions and thus are preferred. Such descriptions can be at the shot level, pre frame or refer to a segment of the video. Shot level 3D quality semantic descriptions include information related to e.g., depth continuity, syncness, or depth stress. Annotation of the VideoSegment representing a shot is done by assigning the 3D quality semantic terms via the How element of StructuralAnnotation. Frame level 3D quality descriptions include information related to e.g., colorimetric or contrast mismatch, and are stored using the same approach as above within the StillRegion that corresponds to the frame. The semantic terms are defined in a respective Classification Scheme file. 3D video quality defects such as stereoscopic window violations, bent window effects or depth jump cuts refer to a segment of the video and the corresponding characterizations are stored as events (see Section III-C). An example of 3D quality description at shot level is given below:

```
<VideoSegment id="vis.shot_0">
 <StructuralUnit href="StructuralUnitCS#vis.shot"/>
 <TextAnnotation>
  <StructuredAnnotation>
   <How href="SemanticFeat3dCS#DisparityMap.high"/>
   <How href="SemanticFeat3dCS#DepthBudget.low"/>
   </StructuredAnnotation>
 </TextAnnotation>
 <MediaTime>
  <MediaRelIncrTimePoint>0</MediaRelIncrTimePoint>
  <MediaIncrDuration>149</MediaIncrDuration>
 </MediaTime>
</VideoSegment>
```

E. Person/Face/Object Description in Stereo Video

The description of an object (e.g., a person, a face, a car, a ball) is performed after object/person detection and tracking that provide the location of an object over time. In the case of humans, additional information can be included, such as depicted activity or facial expression. Semantic characterizations of object/person position and motion can also be stored.

In more detail, object detection is the process of finding the location of a predefined object (e.g. a face, a car, a ball etc) in a per-frame basis [9]. Since the object detector usually detects a specific object or a specific object category, this information can be used to semantically annotate the detected object type. To store the location of the detected object(s), usually in terms of a bounding box, as well as other relevant information, a StillRegion type is used for each detected object, where the specific type (e.g., face, car) is stored, within the StructuralUnit element, while the SpatialLocator type is used to hold the coordinates of the bounding box.

Object/person tracking produces the trajectory of a predefined object (e.g., a face) in a sequence of video frames [10], [11]. Usually, object tracking is initialized by an object detector and performed in a video segment in a frame-by-frame basis [12]. The spatial position of the tracked object (usually in the form of a bounding box) over time forms the object trajectory. The type e.g. face or ball, and the coordinates of the bounding boxes of a trajectory are stored within the StructuralUnit and SpatioTemporalLocator elements of a MovingRegion, respectively.

The resulting StillRegions (detection) and MovingRegions (tracking) are stored within the VideoSegment representing the visual information of the corresponding shot. As can be seen in Figure 1, we use a MediaSourceDecomposition type, to decompose this VideoSegment into a list of StillRegion elements, where each of them represents an entire frame of the video segment. Subsequently, we decompose each frame into further StillRegions representing detected objects through a SpatialDecomposition type. Similarly, we use a SpatioTemporalDecomposition type to decompose the VideoSegment, representing the visual information of the corresponding shot, into a list of MovingRegion elements where each of them represents an object trajectory. In 3D video content, correspondences between StillRegions (e.g. bounding boxes) or MovingRegions (e.g. trajectories), that correspond to the same depiction of the object in two channels, e.g. between the right and left video channel, should be established. To denote such a correspondence, we use the Relation type in order to connect two Segment types, namely StillRegions and MovingRegions across channels.

After object/person detection and tracking, various video analysis algorithms may be used for extracting semantic descriptions for these entities, e.g., information regarding person identity, facial expression or activity and/or object motion. More specifically, human activity recognition can be used to recognize specific predefined human activities, such as run or walk on a person trajectory. Such information can be stored using the WhatAction element of the StructuredAnnottion of the respective StillRegion or MovingRegion. Moreover, facial expression recognition labels, such as happiness, anger, fear etc., can be stored in the WhatAction element. The value “affect” is used within the StructuredAnnotation of the respective StillRegion or MovingRegion, whereas the How element holds the recognized expression.

By using geometrical reasoning algorithms, one can also annotate 3D video content with information related to the geometric or motion properties of objects or object ensembles [13]. Such properties may refer to the geometrical position of an actor or object in the 3D world, the size of an actor/object, the location of a displayed object with respect to the screen or the stereoscopic comfort zone, the direction of object motion etc. In the case of stereo video, object position and motion characterization in the depth domain can be obtained. Such geometric descriptions are stored using appropriate elements of the StructuralAnnotation type of StillRegions and MovingRegions. Specifically, the Where, WhatObject and How elements are used to store labels related to the position (e.g., “left”, “near”, “in front of screen”), size and movement (e.g., “forward”) of an object, respectively.

Image regions depicting faces (facial images) can be clustered into clusters of actors though a facial image clustering algorithm [14]. For storing the results of face/object clustering (i.e., labels such as Actor_1, Actor_2,...), we update appropriately the Who/WhatObject element, respectively, in the structured annotation of each involved segment type (StillRegion and MovingRegion).

It should be noted that in all above cases, the href attribute of the various elements (namely Who, WhatObject, How, Where) of StructuralAnnotation is updated by utilizing appropriate terms from corresponding Classification Schemes.

Relations between faces and real persons may be obtained through manual annotation or through a face recognition algorithm. For example, a face, that is automatically detected and is given an abstract face term (e.g., Actor_1) after face clustering, may be annotated as depicting a certain real person (e.g., Jack Smith). Hence, relations between abstract face terms and real names may be obtained. Each such relation is represented by a StructuredAnnotation which contains two ’Who’ elements: the first ’Who’ element holds the corresponding abstract face term and the second a term from a Classification Scheme containing real names of actors. The resulting StructuredAnnotations are stored in the top level structure for content description, namely the AudioVisual.

In the following XML example, a face trajectory, where the depicted person is happy and running, as well as the relation between Face_1 and Jack Smith is described:

```
<Description xsi:type="ContentEntityType">
 <MultimediaContent xsi:type="AudioVisualType">
  <AudioVisual id="CHANNEL_left" mediaTimeUnit="PT1N25F">
   <StructuralUnit href="StructuralUnitCS#channel.left_view"/>
   <TextAnnotation><StructuredAnnotation>
     <Who href="PersonCS#Jack_Smith"/>
     <Who href="FaceCS#1"/>
   </StructuredAnnotation></TextAnnotation>
   <MediaTime>
    <MediaRelIncrTimePoint>0</MediaRelIncrTimePoint>
    <MediaIncrDuration>1000</MediaIncrDuration>
   </MediaTime>
   <TemporalDecomposition
   criteria="DecompositionCS#shots" id="ShotSet_1">
   ...
       <SpatioTemporalDecomposition
            criteria="DecompositionCS#humans">
        <MovingRegion id="MovingObject_97">
         <StructuralUnit
           href="StructuralUnitCS#face.trajectory"/>
         <TextAnnotation>
          <StructuredAnnotation>
           <WhatAction href="ActionCS#run"/>
          </StructuredAnnotation>
          <StructuredAnnotation>
           <WhatAction href="ActionCS#affect"/>
           <How href="AffectCS#happiness"/>
          </StructuredAnnotation>
          <StructuredAnnotation>
           <Who href="FaceCS#1"/>
          </StructuredAnnotation>
         </TextAnnotation>
         <Relation strength="1.0" target="MovingObject_121"/>
         <SpatioTemporalLocator>
          <ParameterTrajectory motionModel="still">
           <MediaTime>
            <MediaRelIncrTimePoint>188</MediaRelIncrTimePoint>
            <MediaIncrDuration>1</MediaIncrDuration>
           </MediaTime><InitialRegion><Polygon>
              <Coords mpeg7:dim="2 4">38 33 0 -33 19 0 35 0
             </Coords></Polygon></InitialRegion>
          </ParameterTrajectory>
         </SpatioTemporalLocator>
        </MovingRegion>
       </SpatioTemporalDecomposition>
       ...
   </TemporalDecomposition>
  </AudioVisual>
 </MultimediaContent>
</Description>
```

F. Soccer Analysis in Stereo Video

3D Soccer analysis [15] extracts diverse semantic informations, such as camera view class, players positions and identities and interesting highlights [16]. More specifically, the goal of camera view classification for football videos is to recognize different camera views. Usually, four different camera view classes can be observed during a broadcast football match: (i) overview, which is the main camera placed usually on the stands, (ii) close-up, these are views taken from cameras around the field showing a single player. (iii) Medium views are also taken from cameras around the field but show more than one player. (iv) Out-of-field denotes those shots not showing any field action but either audience or the coaching bench. Such class label is stored as a How element within the StructuralAnnotation of the VideoSegment representing the visual information of the respective shot.

Soccer highlights such as corners, goals, etc can also be annotated. This is sensible, as soccer broadcasts usually follow a characteristic cinematic model. For example, most goal scenes start with the shot on the goal shown by an overview camera, followed by medium and/or close-up shots of celebrating players, coaches and audience. Goal scenes typically end with a replay, showing the goal shot from different perspectives. Soccer highlights can be stored using the same approach used for annotating events (see Section III-C).

Finally, players can be detected and tracked for the duration of their appearance. For identifying players, the most obvious way of performing player identification is by recognizing jersey numbers. Similar to other subjects (e.g., actors in a movie), a MovingRegion is used to represent a player's trajectory, while its name is stored in the Who element of the StructuralAnnotation type. In the following example, a shot of a soccer match is characterized as overview shot and a corner is stored as an event:

```
<VideoSegment id="vis.shot_1">
 <StructuralUnit href="StructuralUnitCS#vis.shot"/>
 <TextAnnotation><StructuredAnnotation>
   <How><Name>Overview</Name></How>
  </StructuredAnnotation></TextAnnotation>
 <MediaTime>
  <MediaRelIncrTimePoint>149</MediaRelIncrTimePoint>
  <MediaIncrDuration>91</MediaIncrDuration>
 </MediaTime>
 <TemporalDecomposition
     criteria="DecompositionCS#events">
 <VideoSegment id="EVID_0">
  <StructuralUnit href="event.corner"/>
  <MediaTime>
   <MediaRelIncrTimePoint>250</MediaRelIncrTimePoint>
   <MediaIncrDuration>49</MediaIncrDuration>
   </MediaTime>
  </VideoSegment>
 </TemporalDecomposition>
</VideoSegment>
```

IV. Conclusions

In this paper we have presented a new way of using the AVDP profile of the MPEG-7 standard for 3D video content description that can accommodate multiple audio and video (left/right, color/disparity) channels. We have detailed how we can store the results of several audiovisual analysis algorithms within such a context and how specific 3D metadata can be incorporated in the framework. The proposed framework can be used for storing the analysis results in a 3DTV media asset management (MAM) system. Such a system can then be queried and return the video e.g., where actor X appears in the comfort zone in a 3D context. The proposed XML description can be easily extended to support multiview video coming from multiple cameras.

Acknowledgment

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

References

[1] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Muller, and M. Lang, "Three-dimensional video postproduction and processing," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 607–625, 2011.

[2] B. Mendiburu, *3D Movie Making - Stereoscopic Digital Cinema from Script to Screen.* Focal Press, 2009.

[3] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Language*. J. Wiley, 2002.

[4] I. T. .-A. 1:2012, "Audiovisual description profile (avdp) schema," 2012.

[5] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 82–91, 2006.

[6] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot boundary detection and condensed representation: A review," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 28–37, 2006.

[7] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-based video key frame extraction for low quality video sequences," in *10th Workshop on Image Analysis for Multimedia Interactive Services*, 2009, pp. 25–28.

[8] S. Delis, N. Nikolaidis, and I. Pitas, "Automatic 3d defects identification in stereoscopic videos," in *20th IEEE International Conference on Image Processing*, Sept 2013, pp. 2227–2231.

[9] N. Nikolaidis, M. Krinidis, G. Stamou, and I. Pitas, “Motion tracking in video,” in *The Essential Guide to Video Processing*, A. Bovik, Ed. Elsevier, 2009.

[10] O. Zoidi, A. Tefas, and I. Pitas, “Visual object tracking based on local steering kernels and color histograms,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 5, pp. 870–882, 2013.

[11] H. Baltzakis, A. Argyros, M. Lourakis, and P. Trahanias, “Tracking of human hands and faces through probabilistic fusion of multiple visual cues,” in *Proceedings of the 6th International Conference on Computer Vision Systems*. Springer-Verlag, 2008, pp. 33–42.

[12] O. Zoidi, N. Nikolaidis, and I. Pitas, “Appearance based object tracking in stereo sequences,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 2434–2438.

[13] N. Papanikoloudis, S. Delis, N. Nikolaidis, and I. Pitas, “Semantic description in stereo video content for surveillance applications,” in *Int Workshop on Biometrics and Forensics*, April 2013, pp. 1–4.

[14] G. Orphanidis, N. Nikolaidis, and I. Pitas, “Facial image clustering in 3d video using constrained ncut,” in *21st European Signal Processing Conference*, September 2013.

[15] T. D’Orazio and M. Leo, “A review of vision-based systems for soccer video analysis,” *Pattern Recognition*, vol. 43, no. 8, pp. 2911–2926, 2010.

[16] A. Linnemann, S. Gerke, S. Kriener, and P. Ndjiki-Nya, “Temporally consistent soccer field registration,” in *20th IEEE International Conference on Image Processing*, Sept 2013, pp. 1316–1320.