Anthropocentric Descriptors and Description Schemes for multi-view video content

Ioannis Tsingalis, Nicholas Vretos, Nikos Nikolaidis and Ioannis Pitas
Department of Informatics
Aristotle University Of Thessaloniki
Thessaloniki, Greece 54154
Email: itsingal@csd.auth.gr, { vretos, nikolaid, pitas }@aiia.csd.auth.gr

Abstract—In this paper a new framework for multi-view video content is discussed. The latter framework is based on the MPEG-7 description schemes and is an extension of the Anthropos-7 framework. Furthermore, we propose a new structure that is based on Anthropos-7 and extends the description from single-view to multi-view multimedia content. Moreover, we show that the proposed structure can be used to describe stereo, video plus depth and multi-view video content. The aim of this proposal is to achieve better results in the indexing, filtering and retrieval processes of multi-capturing systems in terms of time complexity.

I. INTRODUCTION

With the advent of the new century a huge amount of information overcame the web. This fact produces new challenges such as the organization, indexing, filtering and finally the retrieval of information. In order to overcome these problems several descriptors and description schemes were introduced. MPEG-7, was the first attempt to tackle with this problem proposing a set of several descriptors and description schemes, the Description Definition Language(DDL) and the description encoding [1]--[4]. However, the large set of these descriptions makes MPEG-7 practically unusable. In order to overcome this problem a new framework was introduced called Anthropos-7. In the latter framework a variety of MPEG-7 based Description Schemes(DS) are used in order to describe the movies content. Until now the proposed descriptors are based on the fact that the whole scene is captured by cameras that are nor geometrically or semantically connected. This means that the camera associated descriptions are nor relevant or semantically connected to each other. This semantic connection becomes important with the new trend of 3D representation of the scene, where the cameras must be connected geometrically or semantically. The connectivity of the cameras is portrayed on the connectivity or the correlation between the semantic descriptions that are obtained from each camera. In this paper we implement this semantic connectivity through the proposed structure.

The remainder of this paper is organized as follows. In Section II, we refer to the previous work on the Anthropocentric model. In Section III, based on the previous work we propose a structure in order to manipulate the Anthropocentric XML descriptions extracted from different capturing cameras. In Section IV, we implement a variety of experiments in order to evaluate the efficiency of the structure. Finally, in Section V, conclusions are discussed.

II. PREVIOUS WORK

In this section we discuss about the basic Description Schemes(DS) of the Anthropos-7 framework in order to provide the bases of the proposed structure. The Description Schemes(DS) are the MovieType Description Scheme, the SceneType Description Scheme, the TakeType and ShotType Description Schemes, the ActorAppearanceType and ObjectAppearanceType Description Schemes and finally the ActorInstanceType and ObjectInstanceType Description Schemes.

A. MovieType Description Scheme(DS)

The MovieType Description Scheme [5][6][7] contains a variety of information about the movie. This kind of information is for instance the title of the movie, the names of the actors and the directors of the movie, the crew and verbal description of the movie. This information could be considered as static information. Of course all these types of information are compatible with the MPEG7 description schemes. Furthermore, the MovieType DS contains a list of scenes, where each scene is represented by an instance of the SceneType. Moreover, MovieType DS can support different versions of the same movie. This is achieved by the VersionType DS, which is a structured collection of scenes from different movie editing processes. The MovieType DS is discussed in details later.

Figure 1, shows the MovieType DS.

B. SceneType Description Scheme(DS)

The SceneType DS [5][6] is introduced in order to organize the movie into hierarchical scene segments. The SceneType DS provides information such as the start and end timecode of the specific scene, as well as information about duration. The Scene Topic tag provides information about the semantic meaning of the scene. The High Level Semantics tag is used in order to describe specific actions on the scene in a semantic way providing a narrative description of the scene. The Sounds [5] tag, apart from the low-level information, also contains high level information for instance who are talking on the scene. Finally, the TakeType and the ShotType DSs are also contained on the SceneType DS and are discussed in details later. Figure 4, shows the SceneType DS.
C. TakeType and ShotType Description Schemes (DSs)

The ShotType DS [5][6] has two versions. One that includes frames and one that does not include frames. In the frameless version, the shot is the basic unit of the movie and it cannot be further divided. Both cases share common attributes such as the ActorAppearanceType and the CameraUseType description schemes. Furthermore, Color and Texture information are also contained in each version.

The TakeType Description Scheme contains both low and high level information. This DS is actually a continuous shot captured by a single camera. This DS has two special abilities, the ability to edit the movie content and the ability to synchronize takes captured by different cameras. The latter functionality is provided by the SynchronizedWith tag and is an element of fundamental importance, especially for the multi-camera based capture systems.

The difference between TakeType and ShotType DS is that Shot instances can not be overlapped temporally while Take instances can. However, in final state of video production, after the postproduction, the Take instances are simplified into shots discarding the overlapped information. Figure 2 and 3 shows the TakeType and the ShotType DSs respectively.

D. ActorAppearanceType and ObjectAppearanceType Description Schemes (DSs)

The ActorAppearanceType DS [5][6] and ObjectAppearanceType DS [5][6] describe the temporal appearance or disappearance of an actor/object on the scene. Also contains low level information about the motion of the actor/object. This information is stored on the Motion tag which is also an MPEG7 compatible scheme. Furthermore, Event tags are included. Finally, the ActorInstancesType DS is also contained on the Actor/Object AppearanceType DS and is discussed latter on. Figure 5, shows ActorAppearanceType DS.

E. ActorInstanceType and ObjectInstanceType Description Schemes (DSs)

The Actor/Object InstanceType DS [5][6] contains low level information about an actor/object within a frame. The BodyPartsType DS is also used and contains information about a specific body part, which is defined by a Region Of Interest (ROI). This ROI is provided by theROIDescription-ModelType DS and can be either a bounding box or a convex hull. Finally, since actors behavior is important, a Status tag is used in order to store information about their facial expressions and/or gestures. Figure 6, shows the ActorInstanceType DS.
one unit, in order to be distinguished from the rest code. This information is placed on the MovieType Description Scheme (DS) and is optional, which means that the whole semantic description is not affected by the absence of the Correspondences tag. See Figure 7 and 1.

B. Corresponding Instances

The CorrespondingInstances tag is the group of the matched instances. The number of the matched instances, depends on the number of the cameras. So in a N-view system the maximum number of the matched instances that are grouped by the CorrespondingInstances tag can be N. The same way, in a stereo or in a video plus depth system the matched instances are two. The last observation indicates that the proposed structure is compatible with stereo, multi-view and video plus depth systems. Finally, an id attribute is also included in order to distinguish the CorrespondingInstances tags. See Figure 7.

C. Corresponding Instance

The CorrespondingInstance tag represents one of the matched instances that are grouped by the CorrespondingInstances tag. This means that the CorrespondingInstance tag contains specific information about the matched instance. This information is the Actor’s/Object’s Instance id and it’s relevant xPath [8]. Both of the attributes are very important and serve specific reasons. The id attribute can be used for an id-based correlation, however this kind of correlation is not always valid because the matched instances do not always have the same id. In order to overcome this problem the xPath attribute is also included and solves the aforementioned ambiguity, as the xPath attribute specifies the exact place of the Actor/Object Instance in the code description. Finally, each CorrespondingInstance tag also contains an identification id which is used in order to distinguish the CorrespondingInstance tags. See Figure 7.

IV. EXPERIMENTAL RESULTS

In this section experimental results of the proposed structure are presented. The XML files we use in the experiments are a combination of two XML files each of them obtained from a left and a right stereo channel. Moreover, additional information of the proposed structure is also included. The first column of Table I refers to the size of each XML file. The second column refers to the number of the Corresponding
Instances and the last column refers to the query’s average execution time. For the experimental implementations XQuery language[9] and the native XML database eXist-db¹ are used. It is obvious from Table I that the query’s average execution time is increasing as the relevant size of the XML file is also increasing. In other words the effectiveness of the proposed structure is related with size of the XML files and consequently with the number of the Corresponding Instances.

<table>
<thead>
<tr>
<th>Size (MB)</th>
<th>Corresponding Instances</th>
<th>Average execution time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>8635</td>
<td>1168</td>
</tr>
<tr>
<td>19.3</td>
<td>6000</td>
<td>468</td>
</tr>
<tr>
<td>13.4</td>
<td>4150</td>
<td>238</td>
</tr>
<tr>
<td>4.83</td>
<td>1500</td>
<td>43</td>
</tr>
<tr>
<td>1.35</td>
<td>421</td>
<td>13</td>
</tr>
</tbody>
</table>

TABLE I: Experimental Results

V. CONCLUSION

In conclusion we underline the fact that the proposed structure can be very useful in the semantic correlation of the XML descriptions that are obtained from different capturing cameras in a N-view capturing system. So this structure can be used in the framework of the new trend of 3DTV and more generally in the framework of 3D video semantic description, in order to achieve efficient results in the retrieval process. Finally, the compatibility of the structure in almost all capturing systems, from the primitive stereo to the N-view systems is one it’s most fundamental attributes.

ACKNOWLEDGMENT

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

¹More information can be found in http://exist.sourceforge.net/