A PROPOSAL FOR A VIDEO MODELING FOR COMPOSING MULTIMEDIA DOCUMENT

CECILE ROISIN[1], TIEN TRAN_THUONG[2] AND LIONEL VILLARD[3]

Opéra Project, research unit Rhône-Alpes of INRIA, Zirst - 655 avenue de l'Europe - 38330 Montbonnot Saint Martin
FRANCE
([1]cecile.roisin@inrialpes.fr, [2]tien.tran_thuong@inrialpes.fr, [3]lionel.villard@inrialpes.fr)

Abstract. This paper contributes to the modeling of audiovisual information with a particular focus on the description needs for the composition of video elements (character, shot, scene, etc.) with other media information (text, sound, image, etc) inside multimedia document. This model has been experimented through an authoring and presentation tool called *VideoMadeus*. The resultants are illustrated with several examples of document where spatio-temporal synchronization of video is required.

Keywords: Video modeling, video content, video element, semantic description, spatial/temporal/spatio-temporal synchronization

1 Introduction

Video is a kind of medium which can carry rich and high-capacity sources of information. It is very efficient in bringing information to various audiences and in many fields such as: entertainment, advertisement, education, etc. or it is used just for storing information. Additionally, the need of using video continuously increases resulting in demands for more and more efficiency, intelligence and convenience in its use [22]. More concretely, for the field of multimedia authoring, there is the need of intelligent descriptions of the video content which make it possible to compose the information of video content with other media objects (text, audio, image…) in multimedia documents.

In general, a video application can be divided into three major steps:

Figure 1. Video application chain

The process begins with analyzing the video data (either automatically or manually). The result of this step is then represented in a predefined format, which is standard and more useful for processing. The last step processes this information

depending on different applications. We can observe that most of the recent research studies either concentrate on the 1st step or on the 3rd step. Therefore video applications are often reduced to 2 steps, the 1st and 3rd step. In these works there is no independence between the information resulting from the analysis and the processing carried out in the application. Moreover with the coarse information obtained from the first step, it is difficult to satisfy the various requirements of using video in the processing step.

Therefore, the problem is now to find out a solution to describe as completely as possible the information obtained in the 1st step. This problem raises recently and is marked by the request for contributing to build up a new standard format to describe audio-visual information of Moving Picture Experts Group (MPEG-7) [21] since October 1998. There are a lot of research studies aiming at finding out a model to describe audio-visual information such as: Dublin core (model for indexing Semantic) [6], INA (AEDI-Audio-visual Event Description Interface) [10], Image & Advanced TV Lab-Columbia Uni. [23], CITEC [11], Etc. With the emergence of MPEG-7, all these researches aim at contributing to build up the future standard for the description of audio-visual data content.

The work described in this paper follows a similar approach as these previously mentioned video modeling activities. Our main contribution comes from the understanding of the specific needs required when integrating video into multimedia documents and consists on a proposal for a video modeling that meets these needs. This model has been experimented through the extension a of prototype authoring tool for multimedia documents called *Madeus* [18]. To prevent any confusion between the previous prototype and the tool described in this paper we have called the new one *VideoMadeus*.

The rest of this paper is organized as follows: section 2 presents some examples that illustrate the needs for a descriptive structure of video in a system of authoring and presentation of multimedia documents. Section 3 describes the main features and the architecture of our system Madeus. Section 4 is an architecture overview of our *VideoMadeus*. Section 5 outlines the main structure of our model to describe video content. Section 6 discusses about some applications of that model that can be realized in our system *VideoMadeus*. In section 7, we present the interface and the functions of the S*tructured video* that helps users to describe easily their video structure. We give in section 8 a brief evaluations of this work through its comparisons on modeling, application and editing aspect with existing works. Finally, the current achievements of our work and some perspectives will be given in the last section.

2 Example of the needs

The examples presented in this article result from our applications under development to illustrate our multimedia system. They are grouped into two types of media composition: temporal synchronization and spatio-temporal synchronization. These examples will be used to illustrate our model and its implementation in the other sections of the paper.

2.1 Slideshow example

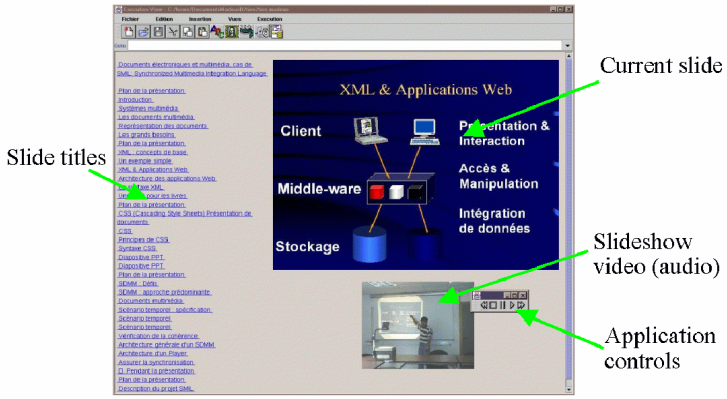


Figure 2. Slide show document example

The Slideshow application is a typical example [1] that illustrates the synchronization needs between the video fragments and the other media objects such as texts and images. In our particular example this is a multimedia document which includes 3 types of media (Figure 2): text objects, to constitute the list of the titles of the slides; image objects, that contain the slides of the presentation; and the video of the presentation of the talk.

What is interesting for the user is to be able to navigate through this talk thanks to synchronization points between these media. A given title allows to access the corresponding slide and the video fragment. In the same way, a given video fragment also allows to access the corresponding slide and the title.

To realize such a document we can see the needs for a video content description: the video has to be decomposed into the fragments corresponding with the different parts of the presentation, as defined by the titles and the slides of the talk and then these video fragments must be synchronized with the text objects and the image objects of the document.

2.2 Spatio-Temporal examples

In the same way, we can find many examples in which spatial synchronization between video objects and external elements are necessary. For instance, the hyperlink on video object (Figure 3); the alignment of text object corresponding to the speech of a character appearing and moving in a video (Figure 4).

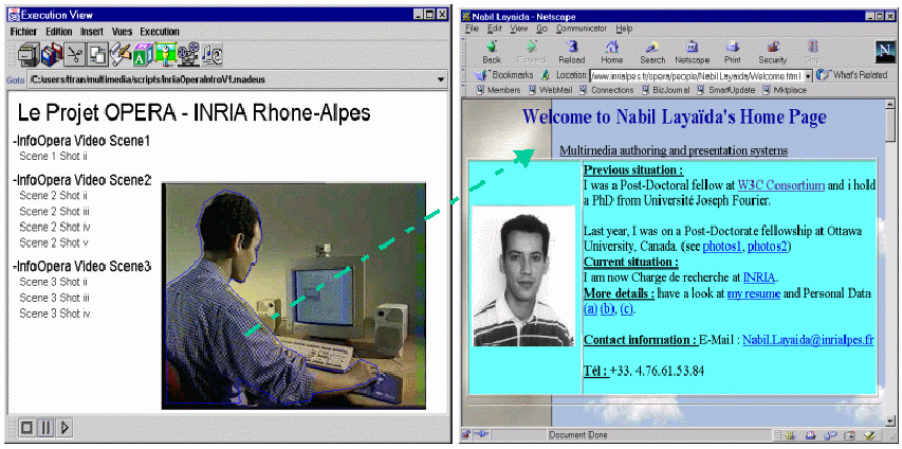


Figure 3. Hyperlink on video object

Video Hyperlink: The example in Figure 3 is an image of the execution view of the multimedia document *InriaOperaIntroduce.madeus*. In this document, the Madeus environment is used to synchronize the video fragments to corresponding text elements on the left. In addition, the application allows the author to define a hyperlink on the occurrence of Nabil character of the video (here to access his home page).

Spatio-Temporal Synchronization: In this example (Figure 4), author has added a textual speech bullet that is presented during the occurrence of a character and that follows his movement in the image. The text elements presenting the speech of character are aligned inside this bullet.

Tracking: Moreover we can define more operations related to the occurrence of the characters in the video like: erase or hide the occurrences of character; enlighten on the occurrence of a character by the addition of a red contour around his drawing; etc. To allow all of these operations we need to identify video objects (their positions, features, etc.) at each time in the video. So a tracking of video object is necessary. Figure 4 illustrates some character tracking applications.

Figure 4. Tracking and saptio-temporal synchronization

The above examples have shown the needs of video decomposition into both temporal and spatial fragments and the corresponding model for describing these fragments. The tools for video decomposition can be partially automatic. They will not be further developed in this paper because we only want to focus on modeling video information. The model allows to describe and locate not only temporally (shot, scene, sequence) and spatio-temporally (occurrence of character) but also semantically (see description semantic example in section 5.3).

3 Madeus, a system of authoring and presentation of document multimedia

Before presenting our video structure model, we rapidly describe in this section the main features of Madeus model. Indeed our video model closely depends on Madeus because it has to be integrated in it to allow media composition as described in the examples of section 2.

3.1 Madeus document model

In Madeus, the description of a multimedia document is organized around four dimensions: logical, temporal, spatial and hypermedia. In this section, we discuss the model for each of these dimensions and show how to combine them together. Its syntax is formally described as a XML DTD and therefore it takes full advantage of all XML existing tools. The DTD itself can be found at [13] and in the remaining part of this article we only use fragments of document instances encoded according to this DTD.

According to the idea of separating document information into dimensions, the general structure of such document instance is decomposed in four main parts:

1. *Content* that describes the content information of the document
2. *Media* that defines how this basic information is used in the document (style information, link, etc.).
3. *Temporal* for the synchronization between document parts
4. *Spatial* for layout specification

Due to its intrinsic nature, the hypermedia information to gather with interactivity is encoded in Temporal and/or Media parts.

A Madeus XML source document look like the examples of figure 11, 12, 13, 14, 15 in section 6. Each part is detailed in the following sections.

3.1.1 Content model

A multimedia presentation is composed of a set of *media*, for instance a picture, a sound, a 3D animation, etc. In order to reuse the same content, its specification is separated from its using context. The *Content* element contains low level information about the media, for instance, pixels of a picture, characters of a text, etc., and the intrinsic properties of the media, like the duration of a video or its size.

3.1.2 Media model

The media part allows to define objects as they will be used really in the multimedia presentation. Basically, an object is defined by a content and style information. The *Content* attribute contains a link to a content description defined in the content part. The description of where and when objects are presented is done respectively in the temporal and spatial parts.

3.1.3 Temporal model

This model allows to organize media objects over time [17]. Every *Media* element is associated with a temporal *Interval* element that carries all the temporal attributes required for its schedule (begin, duration and end). The synchronization is specified both by composite nodes and temporal relations. A composite node enables to temporally group interval elements (see example of figure 12).

3.1.4 Spatial model

The spatial model is basically similar to the temporal model. The main differences are the use of a spatial vocabulary (*left_align*, *bottom_spacing*, etc.) and the extension to support two dimensions unlike the temporal language, which has a single dimension. In addition, a spatial attribute cannot have indefinite value.

More precisely, the spatial model organizes the document space as a 2D box hierarchy. A composite node allows to group set of 2D shapes (*Shape* element) inside 2D boxes (see example of figure 12).

3.2 Madeus architecture

Madeus is based on the Kaomi multimedia toolkit [17], which is a multi-document and a multi-view architecture. The "main" view in which the document is played and various other views conveying comprehending information on the document: its structure, the existing temporal relations and so on. These views can be synchronized on object selection and each one can support editing actions. Finally, a key point is that the author can directly change the document in the presentation view, by stopping the execution of the document and then selecting the objects on which editing actions are performed (for instance, to insert a temporal or a spatial relation). This basic manifold functionality allows an easier authoring task and approaches the WYSIWYG paradigm as provided in editors of static documents.

4 Architecture of VideoMadeus

According to the needs explained above and the extensible capacities of our tool Madeus, we have extended the Madeus authoring environment for video content management: both the document model and processing functions have been completed (Figure 5). This authoring tool allows the user not only to compose the video fragments with other media in a rich way, but also to specify the structure of these fragments either through automatic decomposition or through manual editing operations.

VideoMadeus is composed of the following components (Figure 5):

- The video data management component that handles the video content according to our XML description of video. This component produces the Kaomi internal structures [17] using the XML xerces parser.
- The editing and presentation components is composed of:
  - The video edition view (Figure 15) comprises the multiple filters of the video (structure, execution, navigation, semantic, thesaurus) and the edition functions. It can assist the user to easily modify the descriptions of each video fragment.
  - The document execution view (Figure 2, 3, 4) is the main view of Madeus in which the complete documents are played. This view is also an editing view where the author can directly modify some aspects the documents (for instance spatial relation between objects).

It must be noted that these two views are closely synchronized, i.e., when an element is modified in the edition view, it is updated in the execution view as soon as the execution of the corresponding part occurs.

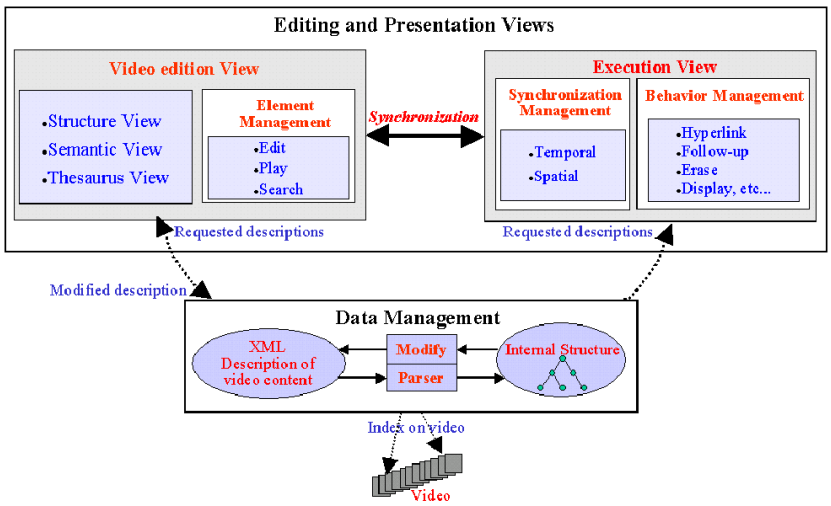


Figure 5. General architecture of *VideoMadeus*

In the rest of this paper we focus on our video content model, then we give some applications of this model for the multimedia edition of documents and finally we describe the video editing view.

5 Model description proposal for video content

In this session we present the principles of our model. We also propose for each presented element a XML specification and show an example of the video (*InriaInfo.mov*) description in the *InriaOperaIntroduce.madeus* document (Figure 3, 4) according to its XML specification. The first version of this model has been presented in [3] and mainly focused on the structure and relations among video components. This preliminary work is now integrated in a more general model presented here, characterized by: three levels of description (see 5.1) and spatio-temporal properties for video objects (see 5.2.2 and 5.2.3)

5.1 General Model

Our model is based on a decomposition of video information into three principal parts: *Structure*, *Semantic* and *Thesaurus* which define three semantic level of knowledge base (Figure 6). In [16] we can find a similar organization of the

knowledge base for a intelligent multimedia system. It is the requirement of classification of knowledge base for the application targeted computation processes, which exploit the knowledge base.

The *Structure* description is a low-level description that directly indexes on raw video to extract the structure of the video. It's the most important part in our model. It makes it possible to describe directly and completely the contents of the video.

At the higher level, there are the *Semantic* description elements, which allow the description of the video contents more semantically (the characters, the events, the relations, etc) (see 5.3).

the *Thesaurus* description elements are the highest level description, which describe semantic terms and expressions to classify elements in the video content description. These terms can be located from a thesaurus or defined by the author (see 5.3).

Other element descriptions: *MetaInfo*, *MediaInfo*, and *Summary* that can make it possible to the author to easily identify video and to look at it quickly and globally. These last three elements are proposed by using concepts from the Dublin core project [6] and from MPEG-7 [21].

In the Figure 6 we can find the general description of the *inriaInfo.mov* video.

```
<VideoContent ID="InriaInfo" FileLocation = "Marion/videos/InriaInfo.mov" … >
    <Structure ID="InriaInfoStruc"  … > . . . </Structure>
    <Semantic ID="InriaInfoSemantic"  … > . . . </Semantic>
    <Thesaurus ID="InriaInfoThesaurus"  … > . . . </Thesaurus>
    <MetaInfo ID="InriaInfoMetaInfo" Author="INRIA" Language="France"  Publication="…" … > … </MetaInfo>
    <MediaInfo ID="InriaInfoMediaInfo" System = "PAL" FileFormat = "MPEG" … > … </MediaInfo>
</VideoContent >
```

Figure 6. General model for the *InriaInfo.mov* video

5.2 Description of the video structure

5.2.1 High level structure

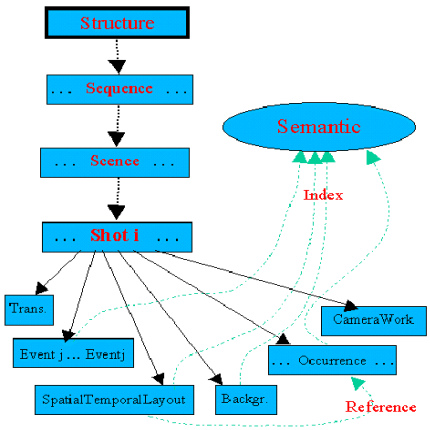
We have defined a model of the video structure very similar to existing works [11, 14]: a video is composed of the successive sequences, a sequence contains successive scenes and a scene contains successive shots (Figure 10). In our work on these levels we have especially focused on the specification of the temporal fragment and relations between them [3].

5.2.2 Shot description

Shot is the smallest unit in classic film theory and defined as the piece of film between two cuts, it is an unbroken take from the start to the switch off of the camera [7]. So in video analyzing, the shot detection is always the first work [15].

Moreover, there is interesting information that can be extracted from a shot according to each application needs: for instance, in a research application we need to extract the feature of the shot as the color histogram, the background, the spatial relations between objects, etc. in a video composition application it is necessary to locate the occurrences of video objects, while in a surveillance system the detection of the moving objects is of high importance, etc. Thus, to complete the information extracted from the analysis phase, our model proposes that a shot is mainly composed of the following elements (Figure 8): *Transition* [7], *Background*, *CameraWork*, *Occurrence* (detailed in section 5.2.3), *Event* and *Spatio-TemporalLayout*.

The *Event* element describes a particular situation in the video that is considered as a relevant part such as a motor explosion, a plane taking off, a demonstration, a storm, etc. It is simply a video fragment in a shot, which is composed, of a sequence of images. It refers to an *EventSemantic* element for a semantic meaning (see section 5.3). The *Spatio-TemporalLayout* element describes the spatio-temporal relationships among the composing regions of occurrences in the shot. There are several different ways to describe, to compare and to retrieve the spatio-temporal relationships, such as spatio-temporal logic [1], augmented transition network (ATN) [24], 2D-strings [25] and its extensions 2D B-strings [26], 2D C-strings [8], etc. and AMOS system of MPEG-7 [20]. The *Spatio-TemporalLayout* description element proposes a XML specification for these methods.




```
<!-- XML Shot Description -->

<!ELEMENT Shot          (Transition?, Background?,
                         Occurrence*, Event*, SpatialTemporalLayout?) >

<!ELEMENT Transition EMPTY >

<!ELEMENT Background  (Region+)>

<!ELEMENT CameraWork  (CameraMotion?) >

<!ELEMENT Occurrence  (Trajectory?,Region+,Occurrence*) >

<!ELEMENT Event       EMPTY>

<!ELEMENT SpatialTemporalLayout   (2D-BStrings?,
                         2D-CStrings?, ATN?, AMOS?, . . . ) >
```

```
<Shot ID=" Shotii"  Start_Time ="5.65" Stop_Time ="8.08" … >
    <Transition ID="TransShotii"  TypeTransition="fade-in"… />
    <Background ID=" BackgroundShotii"  ColorHistogram="rgb(88,84,45)"… > …  </Background>
    <CameraWork ID=" CameraShotii"   … > … </CameraWork>
    <SpatialTemporalLayout ID=" SpatiallayoutShotii"  …  > … </SpatialLayout>
    <Event ID=" NabilMoveMouse"   Start_Time ="6.5"  Stop_Time ="6.9" . . . > . . .</Event>
    <Occurence ID="OccNabil1" Start_Time ="5.621" Stop_Time ="7.021"… > … </Occurrence>
</Shot>
```

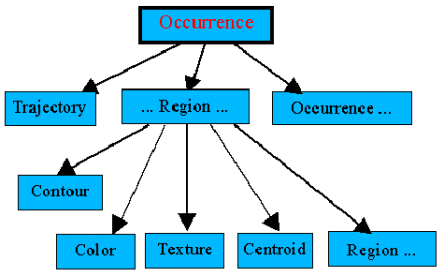
Figure 7. Shot description.

5.2.3 Occurrence description

The *Occurrence* element describes a character or an object that appears in the shot. It refers to a video object element in the semantic description that corresponds to that character or object (see section 5.3). Description of the occurrences enables us to associate actions to the appearances of video objects like hyperlink, filter, seek, follow, synchronize spatio-temporal with other media, etc.

In our model, an occurrence description is composed of (Figure 9):

- A *Trajectory* description [4], which is the description of the movement of significant points on the occurrence: its central point, the eyes of a character, etc.
- A list of *Regions* key descriptions that describe the occurrence at particular positions in time. This will allow to implement tracking and searching objects [24]. Each region can have element descriptions, which define properties of the region, such as contour, color, texture, centroid and its included regions. Interpolators use these properties to determine intermediate regions. Recently, we have used the spline interpolation [9] and polygon interpolation [19] for the location of the occurrence at each time.
- And it is possibly to have other occurrence elements inside the occurrence like components of the object [23], for instance, the arms of a character, his clothing, etc.




```
<!-- XML Occurrence description -->

<!ELEMENT Occurrence    (Trajectory*, Region+,
                         Occurrence*) >

<!ELEMENT Trajectory    ... >

<!ELEMENT Region        (Contour, Color?, . . .) >
    <!ELEMENT Contour   (InstantContour+)>
    <!ELEMENT Color     ...>
    <!ELEMENT Texture   ...>
    <!ELEMENT Centroid  ...>
```

```
<Shot ID="Shotii" Start_Time ="5.65" Stop_Time ="8.08" . . . >
    <Occurrence ID="OccNabil1" Start_Time ="5.621" Stop_Time ="7.021" … >
      <ListKeyRegion> . . .
        <Region ID=”Reg1” … > . . .
          <Contour ID=”Contour1” KeyTime=”5.62” … >-190,286 -193,254 -125,287 … </Contour>
        </Region> . . .
        <Region ID=”Reg4” … > . . .
```

```
            <Contour ID="Contour 2" KeyTime="7.02" …> 114,281 111,249 179,282  …</Contour>
          </Region>
        </ListKeyRegion>
      </Occurrence>
    </Shot>
```

Figure 8. Occurrence description

The example of the Figure 8 is the description of the *OccNabil2* occurrence in the scene 3, shot 4 of our *InriaInfo.mov* video.

5.3 Element Semantic

We have enriched our description video content model with *Thesaurus* and *Semantic* elements, which allow semantic query and semantic search operations on the video content [27][28]. In this paper we do not develop our semantic model for video, but we want to focus on the importance of its integration inside the global video structure thanks to XML link capabilities. For instance, Figure 10 is a part of a description of the video Lion King. In this description, the occurrences of Simba are grouped in a video object Simba by references to the video object Simba towards its occurrences and the reverse. At the more semantic level, the video object Simba is referred by an object Lion in the thesaurus description, which informs that the Simba character is a lion.

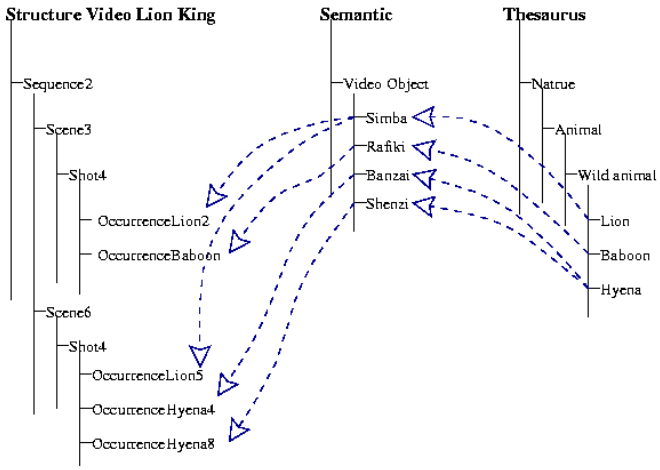


Figure 9. Example of a semantic description

6 Application of the model in Madeus

The video content description model above is used to compose the video content element with other elements (text, image, sound, etc.) in the Madeus authoring environment. In this section we describe step by step the composition of the document *Madeus* (InriaOperaIntroduce.madeus) by using the video structured *InriaInfo.mov*.

6.1 Video content description in document multimedia Madeus

Video content description points out the low-level data of a video media. Therefore it can be defined as a content description *VideoContent* in the Content part of the document Madeus (Figure 10).

```
<Content>
    <VideoContent ID="InriaInfo" … >
        <Structure ID="InriaInfoStruc"  … >
            <Sequence ID="SeqOperaInfo" Start_Time ="0" Stop_Time ="41.89"  … > . . .
                <Scene ID="Scene2" Start_Time ="4.17" Stop_Time ="10.16" … > . . .
                    <Shot ID="Shotii" Start_Time ="8.71" Stop_Time ="11.09"  … >  …  </Shot>
                </Scene>
                <Scene ID="Scene3" Start_Time ="10.17" Stop_Time ="28.04" … > …
                    <Shot ID="Shotiv" Start_Time ="15.13" Stop_Time ="17.96"  … >  …  </Shot>
                </Scene>
            </Sequence>
            <Sequence ID="Sequnce2" Start_Time ="41.90" Stop_Time ="76.69" … > … </Sequence>
        </Structure> …
    </VideoContent>  …
</Content>
```

Figure 10. *InriaInfo.mov* video description in the *Content* part of the *InriaOperaIntroduce.madeus* document

This example shows how the high level structures of the *InriaInfo.mov* video is integrated inside the content part of the document.

6.2 Use of a fragment video like a object media

From the video fragments described in the *VideoContent* part, the author can create instances of these fragments in order to associate presentation attributes (with the *VideoElement* in the Media part), temporal behavior (with the interval element) and spatial properties (with the region element).

```
<Madeus …>
    <Content>  …
        <VideoContent ID="InriaInfo" >  …
            <Sequence  ID = «SeqOperaInfo»  ... > …  </ Sequence> …
        </VideoContent> …
    </Content>
    <Media> …
        <VideoElement  ID=«OperaMemberVideo» Content = «InriaInfo.InriaInfoStruc.SeqOperaInfo»
```

```
                    TypeRenderer = "LightWeight"  …  >  …  </ VideoElement >  …
    </Media>
    <Temporal>
        <Inetrval  ID = «OperaMemberVideoInterval»  Media = «OperaMemberVideo» … />   …
     </Temporal>
    <Spatial>
        <Region  ID = «OperaMemberVideoRegion»  Media = «OperaMemberVideo»  Top= «10» Left=«20» … />
          …
    </Spatial>
</Madeus>
```

Figure 11. Use a fragment video described like an object media.

The example of the Figure 11 shows the definition of the *VideoElement* "OperaMemberVideo" media in the *Media, Temporal* and *Spatial* parts of the document. This media object is defined like an instance of the video fragment description *SeqOperaInfo*.

6.3 Actions on occurrence of video object

An *Occurrence* is a spatio-temporal element that cannot appear as a specific media in the Media part of a document (the smallest entity is an image). Therefore an occurrence only appears as an XML content description inside the video fragment descriptions (event, shot, scene, etc) in the *Content* part. But its description is useful for application of a set of actions on occurrences: hyperlink, tracking and erasing.

In the example of the Figure 12, the *OperaMemberVideo* media references the *SeqOperaInfo* sequence which contains occurrences: *OccIreneVatton1* and *OccNabil1* in scene 2, shot 2 (*Scene2.Shotii*) and *OccNabil2* in scene 3, shot 4 (*Scene3.Shotiv*). Actions are specified in the *OperaMemberVideo* media: the erasing action on the *OccIreneVatton1* occurrence; the tracking action on the *OccNabil1* occurrence and the Hyperlink action on the *OccNabil2* that allows the user to click on the Nabil occurrence of scene 3, shot 4, etc (see figure 2,3,4).

```
<Media …>
    <VideoElement ID="OperaMemberVideo" Content="InriaInfo.InriaInfoStruc.SeqOperaInfo"
            TypeRenderer = "LightWeight"   … >
              <Erase Object = "Scne2.Shotii.OccIreneVatton1"  FillColor="rgb(84,84,44)"  … />
              <Tracking Object="Scne2.Shotii.OccNabil1"  …  />
              <HyperLink Object = "Scene3.Shotiv.OccNabil2"
                   HRef = "file:///C:/Users/ttran/Multimedia/Madeus/StructuredVideo/opera.html"  …  />  …
    </VideoElement> …
</Media>
```

Figure 12. Actions on occurrences of video object

6.4 Temporal synchronization

Presentation scenario of Madeus document is created by connecting intervals of media object using temporal relations of Allen (meets, starts, equals, during,

overlaps, etc.) [12]. In the same way, the intervals of *VideoElement* media object can be related with other media. Moreover the author can put in relation fragments of these intervals of *VideoElement* media.

The example of Figure 13 shows a *Start* relation between video media object *OperaMemberVideoInterval* and the title text *txtTitleOpera*, i.e., in the presentation of the document, the beginning of *OperaMemberVideoInterval* will be synchronized with the beginning of *txtTitleOpera*. In the same way, the author can synchronize the scenes and the shots in the *OperaMemberVideoInterval* media: the beginning of the *Scene2* with the beginning of the *txtScene2* text by the Start relation; the end the *Scene3* and the beginning of the *txtScene4* text by the Meets relation; the beginning and the end of the *Scene3.Shotiv* with the *txtScene3Shotiv* text by the Equals relation (see Figure 4).

```
<Madeus Name="DocMadeus" Version="2.0" Width="800" Height="600">  …
   <Temporal>  …
            <Relations> …
                 <Starts Interval1=   «OperaMemberVideoInterval»  Interval2= «txtTitleOpera» />
                 <Starts Interval1="OperaMemberVideoInterval.Scene2" Interval2="txtScene2" />
                 <Meets Interval1="OperaMemberVideoInterval.Scene3" Interval2="txtScene4" />
                 <Equals Interval1="OperaMemberVideoInterval.Scene3.Shotiv" Interval2="txtScene3Shotiv" />
            </Relations>
   </Temporal> …
</Madeus>
```

Figure 13. Temporal synchronization

6.5 Spatio-temporal synchronization

In the same way, the spatial layout of the Madeus document uses spatial relationships: *left_align*, *right_align*, *center_v*, *center_h*, *signal_align*, *bottom_align*, etc. In particular, if an area of *VideoElement* media object contains occurrences, the author can carry out spatial synchronization between these occurrences with other areas of the other media.

In the example below (Figure 14), the author has aligned the *OperaMemberVideoRegion* area, where the *OperaMemberVideo* media will appear, with the area of the *textTitleOperaRegion* media text by the spatial relation *Top_align*. And in particular, the author has aligned the occurrence *OccNabil1* in the *SeqInfoOpera.Scene3.Shotii* shot with the *textHelloRegion* media text by the spatial relation *Center_v* (Center vertical), see Figure 4.

```
<Madeus … > …
     <Spatial> …
          <Relation>
               <Top_align  Region1=" MovieInriaGen"  Region2="textTitre"/>
               <Center_v  Region1=" MovieInriaGen.Seq.Scene2.Shotii.Occj"  Region2="textOcc."/> …
          </Relation>
     </Spatial>
</Madeus>
```

Figure 14. Spatio-temporal synchronization

7 View of the structured video

As already mentioned, a video application is often composed of three stages and this paper focuses on the second stage: the video description. The problem is not only to find a format for the description of video information, but also to transform information extracted from the video raw into new formats according to the suggested model. Moreover in the description model, there are semantic elements (scene, character, spatial/personal relation, etc.) which currently (or forever) cannot be automatically generated from coarse information by functions of transformation and deduction. Hence it is necessary to provide an environment for helping the author to manually describe the semantic elements. In our search framework, we develop also an environment for that purpose: it is called the *video edition view.*

7.1 Requirements of the video edition view

The video edition view requires: the visualization of video descriptions, the integration of video analysis tools and the edition services such as splitting a scene, specifying semantic elements and relationships, etc.

7.2 Video edition view of VideoMadeus

Currently we are developing in Madeus a prototype of the video edition view. That view integrates: the video segmentation tools to automatically cut out shots in the video; the semi automatic function to extract the occurrence of the video objects; the manual functions to extract the semantic fragments: event, spatial layout, etc. to group shots in scene, scenes in sequence, occurrences in video objet, etc.

Our video edition view is composed of 4 views: *structure view*, *video player view*, *video information view* and *attribute view* (Figure 15). The *structure view* lays out the tree structure of the video content description. It also allows the author to navigate on each description element of the video description. The other views are synchronized with this view in order to allow direct access to the elements selected in the structure view. The *video player view* allows to play a video fragment corresponding to the selected element description in the structure view. The *video information view* displays the current information of the video player, such as the default and real frame rate, the current video time, the current frame. The *attribute view* allows to display and modify the attributes of the selected element in the structure view.

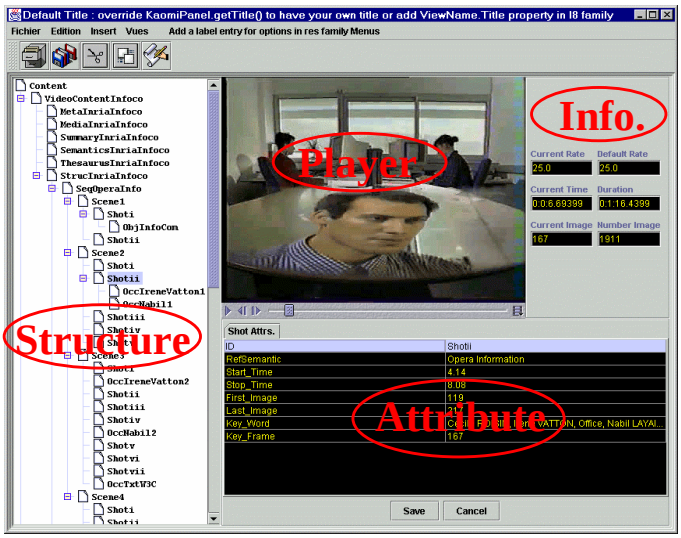


Figure 15. *VideoMadeus* edition view

In addition, the *VideoMadeus* view is synchronized with the hierarchical view of Madeus and therefore can help the author to compose a Madeus document, which contains structured video.

8 Evaluation of the work

Our work provides support for deeper access into video data in our multimedia authoring environment, which until now has treated video data as a black-box, basic medium. In this section, we evaluate this work by comparison with other research in three areas: modeling, application and editing.

Our model is focused on authoring and rendering multimedia documents. It is not designed for searching, indexing or archiving. For this reason, it makes little use of metadata descriptions such as AEDI [10] or MPEG-7 [20,21,22]. Instead, our model is focused on the structural organization of video descriptions that are relevant for media composition. The high level descriptions are very similar to those used in existing work: Sequences, Scenes, and Shots [6,11,14]. We have also introduced lower level descriptions in order to specify details such as Transitions, Events, and Occurrences (see section 5.2.2). In particular, we have proposed the use of spatio-temporal descriptions, which make it possible to place video objects in time. This spatio-temporal description is very important for developing interactive operations on video objects such as tracking, hyperlink, and erasure. In addition, we have also proposed semantic descriptions like those in [27][28], but

have divided them into two levels: Semantic and Thesaurus. This facilitates semantic references to the video elements (see example in 5.3).

The second key point of this work results from its application context. The fact that the model has been fully integrated into our MADEUS multimedia authoring environment verifies that our model is effective in the multimedia document domain. Users of MADEUS can synchronize video objects with other media objects in both time and space and can also apply operations and interactions on elements of the video objects such as tracking, hyperlink, and erasure (see sections 2 and 6). Thus, the author can specify more complex multimedia documents while maintaining the declarative approach of XML that allows the use of high-level authoring interfaces like the one described in section 7.

The last interesting feature is our support for the specification of video descriptions. Our video editing view helps the user create and modify descriptions of structured video data in accordance with our video description model. This view is similar to The IBM MPEG-7 Visual Annotation Tool [29], which is used for authoring MPEG-7 descriptions based on the MPEG-7 Standard Multimedia Description Schemes (MDS). But unlike MDS, our structured video editing interface is not isolated. It is an extension of our KAOMI toolkit, which means that it is synchronized with other views in the MADEUS system (timeline, execution, hierarchy, etc.) (see sections 4 and 7). Moreover, the architecture provides a simple way to integrate existing tools, such as automatic video analysis.

9 Conclusion and perspectives

In this paper we have proposed a model of video description for multimedia applications which can handle video media more finely. It is characterized by temporal/spatial synchronization, actions on video elements (hyperlink, erasing object, tracking an object, etc.) and semantic classification on knowledge source description. We have also described an experimental development of a view helping the author to edit the description of the video and to compose video elements with other media inside a real multimedia document.

For us it is of high importance to cover both a modeling activity and an experimenting activity in the video area. We intend to go further in both directions. Our perspectives in video modeling will refine spatio-temporal relations that are necessary to perform tracking more efficiency. We have also a very simple framework for knowledge specification that requires to be extended because it is central for query on video. With our first experiences, we have been convinced that it is fruitful to integrate research/query services in authoring environments. Therefore we need to have more rich semantic video description.

References

1. A. D. Bimbo, E. Vicario and D. Zingoni, "Symbolic description and visual querying of image sequences using spatio-temporal logic", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No. 4, August, 1995.
2. B.David, G.Anoop, S.Elizabeth and L.Francis, Asynchronous Collaboration around Multimedia and its Application to On-Demand Training. Collaboration and Multimedia System Group Microsoft Research, Redmond, WA 98052.
3. C.Roisin, T.Tran Thuong, L.Villard, ``Integration of structured video in a multimedia authoring system", *Proc. of the Eurographics Multimedia'99 Workshop, Springer Computer Science*, ed., pp. 133-142, Milan, September 1999.
4. C.W.Chang and S.Y.Lee, A Video Information System for Sport Motion Analysis. Journal of Visual Languages and Computing (1997) 8, 265-287.
5. D.Chandler, The 'Grammar' of Television and Film, UWA, 1994, http://www.aber.ac.uk/~dgc/gramtv.html.
6. Dublin Core Metadata Element Set, http://purl.oclc.org/dc/documents/rec-dces-19990702.htm.
7. FILM BASICS: Learning to “Read” & Write about Film, http://www.cocc.edu/cagatucci/classes/wr316/assignments/filmbasics.htm.
8. Fang-Jung Hsu, Suh-Yin Lee and Bao-Shuh Lin, Video Data Indexing by 2D C-Trees. *Journal of Visual Languages and Computing* (1998) 3, pp. 375-397.
9. Foley, vanDam, Feiner, Hughes, Computer Graphics: Principles and Practice. Second Edition, 1990.
10. G.Auffret, J.Carrive, O.Chevet,T.Dechilly, R.Ronfard, B.Bachimont, Audiovisual Event Description Interface AEDI v1.0. User guide, INA, France, 1998.
11. J.Hunter, A Proposal for an MPEG-7 Description Definition Language (DDL), CITEC, Autralia, 1999.
12. J. F. Allen, Maintaining Knowledge about Temporal intervals, CACM, 26 (11), pp. 832-843, 1983.
13. L.Villard, Madeus model DTD, http://www.inrialpes.fr/opera/madeusmodel.dtd, 2000.
14. M. Corridoni Jacopo, D.B. Alberto, D. Lucarella and He Wenxue, Multi-perspective Navigation of Movies. *Journal of Visual Languages and Computing* (1996) 7, 445-466.
15. M. Gelgon, P. Bouthemy, G. Fabrice (1997) A Unified Approach to Shot Change Detection and Camera Motion Characterization; Research report RR-3304 INRIA Rennes, http://www.inria.fr/RRRT/RR-3304.html.
16. Bordegoni M., et al, “A Standard Reference Model for intelligent Multimedia Presentation Systems,” April 1997, pre-print. <http://www.dfki.uni-sb.de/~rist/csi97/csi97.html>.

17. M. Jourdan, C. Roisin, L. Tardif, ``A Scalable Toolkit for Designing Multimedia Authoring Environments", numéro spécial `Multimedia Authoring and Presentation: Strategies, Tools, and Experiences' de *Multimedia Tools and Applications Journal*, Kluwer Academic Publishers, 1999.
18. MADEUS, An authoring environment for multimedia documents. http://www.inrialpes.fr/opera/Madeus.en.html
19. Mark Owen and Philip Willis, Modelling and Interpolating Cartoon Characters. Presented at Computer '94, *Geneva Conference Proceedings* (IEEE May 1994) pp. 148-155; color plates p. 203.
20. MPEG-7 Multimedia Description Schemes XM (Version 2.0), http://archive.dstc.edu.au/mpeg7-ddl/, 2000.
21. MPEG-7 Multimedia Description Schemes WD, *MPEG-7 Description Definition Language (DDL) Home Page*, http://archive.dstc.edu.au/mpeg7-ddl/issues.html.
22. Requirements Groups (Adam Lindsay, Editor), MPEG-7 Applications Document. ISO/IEC JTC1/SC29/WG11/N2861 July 1999/Vancouver. http://www.darmstadt.gmd.de/mobile/MPEG7/Documents/W2860.htm.
23. S.Paek, A.B.Benitez, and S.K.Chang, Self-Describing Schemes for Interoperable MPEG-7 Multimedia Content Descriptions. Image & Advanced TV Lab, Department of Electrical Engineering Columbia University, USA, 1999.
24. Shu-Ching Chen, Mei-Ling Shyu, and R. L. Kashyap, "Augmented Transition Network as a Semantic Model for Video Data," *International Journal of Networking and Information Systems*, Special Issue on Video Data.
25. S.-K. Chang, Q. Y. Shi, and C. Y. Yan, "Iconic indexing by 2-D strings", *IEEE Trans. Pattern Anal. Machine Intell*, 9(3): 413-428, May 1987.
26. S. Lee, M. Yang & J. Chen, Signature File as a Spatial Filter for Iconic Image database. *Journal of Visual Languages and Computing* (1992) 3, pp. 373-397.
27. Shermann Sze-Man Chan and Qing Li, Developing an object-oriented Video Database System with Spatio-Temporal Reasoning Capabilities. Department of Computer Science, City University of Hong Kong.
28. Jia-Ling Koh, Chin-Sung Lee, Arbee L. P. Chen, Semantic Video Model for Content-based Retrieval, IEEE Multimedia Systems '99, Volume 2, p.472-478.
29. Blaise Lugeon and John R. Smith, MPEG-7 Visual Authoring Tool, IBM T. J. Watson Research Center, http://www.alphaworks.ibm.com/tech/mpeg-7.