Annotations as Multiple Perspectives of Video Content

Miguel Costa
4VDO
Rua João Crisóstomo, 49 4 Dto
1000 Lisbon, Portugal
+351 21 3150383
miguel.costa@4vdo.pt

Nuno Correia
DI/FCT/UNL
Quinta da Torre
2829-516 Caparica, Portugal
+351 21 2948536
nmc@di.fct.unl.pt

Nuno Guimarães
DI/FC/UL
Campo Grande, Edifício C5
1749-016 Lisbon, Portugal
+351 21 7500084
nmg@di.fc.ul.pt

ABSTRACT
This paper describes a video annotation tool based on a new and flexible model, that gives several perspectives over the same video content. The model was designed in a way that allows having multiple views over the same video data, enabling users with different requirements to have the most appropriate interface. These views, video-lenses, highlight a specific aspect of the video content that is being annotated. Annotations are made using a timeline based interface with multiple tracks, where each track corresponds to a given video-lens. The format used to store and exchange the information is the MPEG-7 standard. The annotation tool (VAnnotator) is being developed in the scope of Vizard, an ambitious project that aims to define a new paradigm for video navigation, annotation, editing and retrieval. The Vizard project includes users, both from the production/archiving area and from the consumer electronics area, that help to define and validate the annotation requirements and functionality.

Categories and Subject Descriptors
H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – video; D.2.2 [**Software Engineering**]: Design Tools and Techniques – User interfaces; H.3.0 [**Information Storage and Retrieval**]: General

General Terms
Documentation, Design, Experimentation, Human Factors, Standardization, Languages

Keywords
Video annotation; Timeline model; Authoring paradigms; MPEG-7; Video-lens

Multimedia'02, December 1-6, 2002, Juan-les-Pins, France.

1. INTRODUCTION
With the increasing availability of tools to produce digital video content, video archives are becoming larger and more difficult to manage. This is true both for home users and professional users working in the broadcast or production areas. Annotation or metadata provides a way to describe this information in order to improve the management of archives and the retrieval of a given video sequence. There are different approaches for video annotation, reviewed in more detail in the next section. These approaches range from keyword annotation to complex visual languages [5,6]. The need to have a common metadata format led to the development of the MPEG-7 standard [10,11] that is also used in the tools described in this paper. This standard defines a set of descriptors of audiovisual data, a way of describing new descriptors and data structures for indexing and searching by content.

Video information is usually very rich and different users are interested in different aspect of it. The MPEG-7 standard provides a huge amount of descriptors to cover most of these aspects. As such it is a great challenge to provide users with an interface that can be tailored to a specific application. Time associated to multimedia content offers a number of complex questions regarding the presentation of data, such as how to effectively display all the information or how to allow users an easy navigation. In our proposal, we address these questions, by including the representation of time-related content following the timeline model given in Section 3.

One of the goals of our approach is to adapt to the requirements of various kinds of users, providing dedicated perspectives of the data. Also, the collaborative aspect of multi-user working environments is considered. Aimed at providing solutions for these issues, an adaptive interface mechanism, dubbed video-lens, is introduced in Section 4. Our concepts are currently being put forth in the form of a video annotation tool, part of the video editing software under development by the Vizard consortium [2]. Details of this work are presented in section 5.

The next section briefly describes related work. Section 3 presents the model and concepts for timeline based video annotation. Section 4 presents the Video-lens concept. Section 5 describes the tool that implements the models introduced in the previous two sections. The paper ends with some preliminary conclusions and directions for future work.

2. RELATED WORK

There are several approaches to video annotation for different purposes such as improving archival and retrieval or adding contextual information. In [3] there is a review of a set of early systems. One of these systems, the VideoNoter allowed users to create annotations and verbal transcriptions that were used to index a tape or disk. The EVA system, an interactive video annotation tool, had mechanisms for on-line real time and off-line annotations. In this system text transcriptions could appear as "subtitles" synchronized with the video. The Virtual VCR, developed by Buxton and Moran, presents a graphical image of a VCR control panel. Using this control panel it is possible to mark the tape with indices and associated comments. The Vanna system supports a variety of input devices to handle real-time annotation and detailed analysis of video data. MediaStreams [5,6] is a video annotation tool based on a rich visual language that aims to describe all aspects of video content. Regarding more recent systems, Weborama [8] is a multimedia annotation system, describing a client-server architecture for asynchronous annotation sharing, but the emphasis is not on user interface issues and annotation modes for video information. The work reported in [13] is a prototype of a collaborative video annotation tool. It provides a Web based interface for adding and viewing annotations and it allows users to share their annotations through a server. The X-Database [4] is an MPEG-7/XML based system for a collaborative video annotation system. There are several applications that are used to access and manage the database and to annotate the videos locally or using the Web. The database maps the MPEG-7 descriptors and it is automatically created from the XML-Schema. In [1] several tools for automated and semi-automated real time annotation are presented. The metadata (based on MPEG-7) is used for retrieval but can also be used during the different phases of the production process.

Our previous work on annotation systems [7] outlines the principles behind video annotation for adding content or structure, in a way that has similarities with static and printed media annotation. In this sense, the dominant role of annotation is not to describe the material for later retrieval, but a way to augment the documents. The principles are illustrated by a system for video annotation and browsing, named AntV (Annotations in Video). This system allows previewing video streams, creating and editing annotations, and viewing annotated videos. It can import annotations from multiple authors, allowing to share comments about a given video program. As a result of the annotation process, it can create new videos or hyperdocuments based on the original video.

3. THE TIMELINE MODEL

We propose the timeline model of Figure 1 as support for the representation of time-related multimedia content. This kind of representation has a visual equivalent, similar to that used by popular multimedia tools like Adobe Premiere or Macromedia Director, and thus users will most likely be familiar with it. The following description of the timeline model can also be adapted to other time-related applications, e.g. authoring tools or project planning tools.

A timeline is a set of events occurring within a specified duration. Events are organized into tracks, with each event allocated to a given segment of its track. Segments are exclusive within each track, meaning that the intersection of events in the same track cannot occur. For two events to overlap, they must be in different tracks. Events can have a number of tagged values acting as qualifiers. A tagged value is simply an instance of some data-type, labeled with a string of text. The model does not specify the semantics of this qualification. In our framework, video-lenses provide this semantics by defining interpretations (visual and otherwise) of events and their qualifying tagged values. Video-lenses are introduced in Section 4.

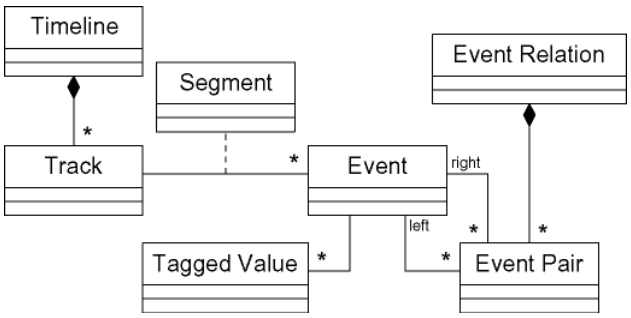

Figure 1: The Timeline model

The model provides also the means to define arbitrary relations between events by allowing the definition of sets of event pairs. These can be used to establish structure among the events, other than the implicit linear structure induced by the assignment of events to tracks and by the sequential order of events within each track. For example, in the case of a project planning application, with timeline events representing scheduled tasks, event relations can be useful for defining sub-tasks of tasks. In the case of an audiovisual management tool, such as the video annotation application presented here, described event relations could be used to establish a hierarchy between events representing single camera shots and events representing scenes that encompass those shots.

4. VIDEO LENSES

Video-lenses provide interpretations of the multimedia content, giving different perspectives of the information, according to the specific requirements of each kind of users. Thus, the same piece of information can be visualized and modified in different ways according to the video-lens being used. The video-lens concept was inspired by the information visualization tools developed at Xerox Parc [12].

In the example of audiovisual content enhanced with information of scenes, one user may be interested in technical information of a scene, whereas another is interested in the characters and script information for that same scene. The one can resort to a video-lens that allows technical information to be entered and viewed for a scene. The other can use a second video-lens that also allows entering and viewing data, for the kind of information that he is interested in. Someone supervising the work of the two previous users could use a third video-lens that provides a summary of both technical and non-technical information.

The interpretation of video-lens can also result in other non-visual side effects. For example, the usual interpretation of an audiovisual data stream would be to display the images that correspond to the video information and to produce sounds that

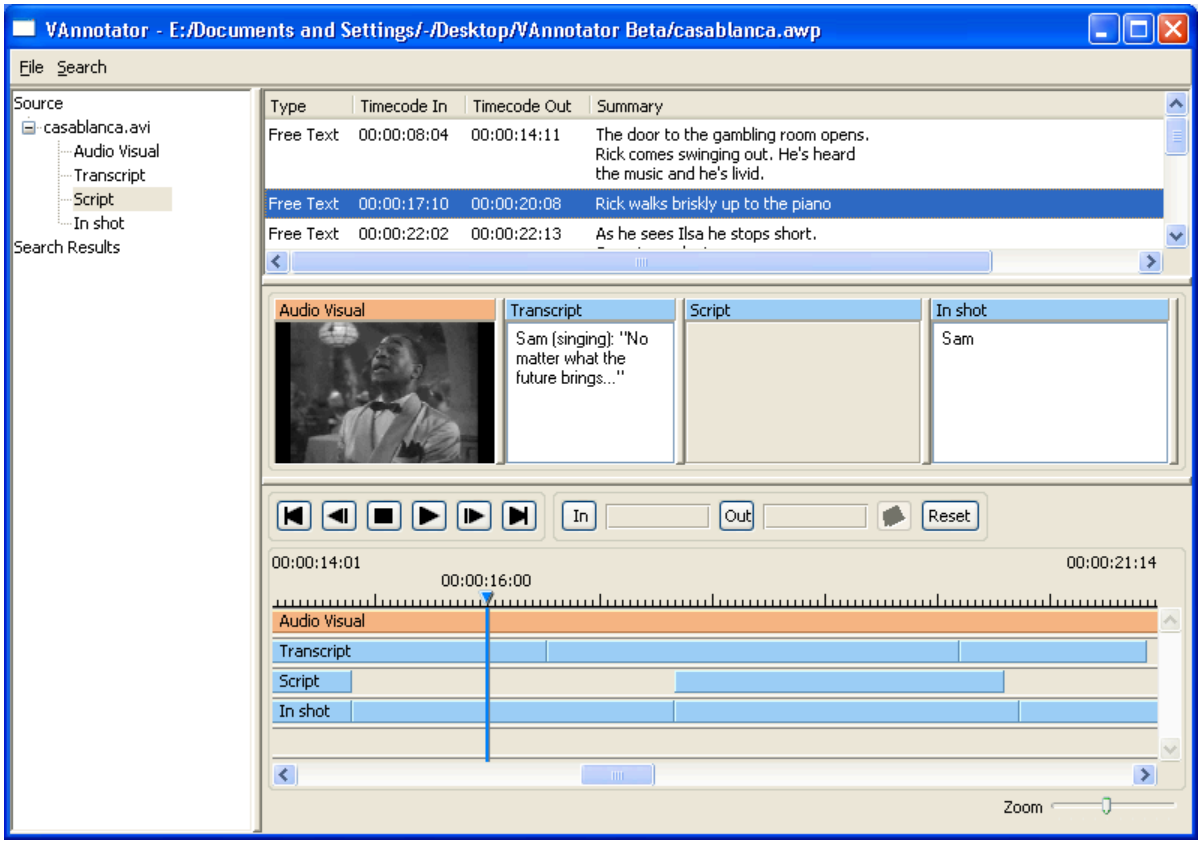

Figure 2: The VAnnotator application

correspond to the audio information. However, we could apply a video-lens to an audiovisual data stream, whose interpretation would produce a list of cuts or calculate dominant color information.

In the context of a timeline following the model given in Section 3, video-lenses provide interpretations of the various events, including their associated tagged values. Video-lenses can use these values to provide data fields for users to edit. Again in the example of audiovisual management, a user can create an event in the timeline, intending to use it as a description of a scene, and then associate this semantics by applying a scene description video-lens. The interpretation of this event by the video-lens would be to display a form that could be used to enter information for the scene.

Initially, no tagged values are present when an event is created. A video-lens can create the tagged values that it requires to provide its interpretation of an event. In the previous case of scene descriptions, once the video-lens is applied to a newly created event, it adds tagged values corresponding to the fields in the scene description form that is displayed to the user.

5. VIZARD VANNOTATOR

The Vizard project [2] is currently developing an advanced video publishing tool based on new concepts of how to deal with video content. The Vizard system consists of three modules: (1) VExplorer, a video collection search, organization and management tool; (2) VPublisher, a new-generation storyboard, video editing and video publishing tool; and (3) VAnnotator, a flexible and intuitive video annotation tool, based on the approach to multimedia management presented herein.

VAnnotator allows users to annotate audiovisual content by providing a timeline where events that correspond to annotations can be created. Each track of the timeline has an associated video-lens, whose user interface is also be available next to the timeline. Users are able to enter annotations by means of these video-lenses and later view those annotations during playback. Creating new types of annotations will also be possible by adding new definitions of video-lenses.

The current interface of the VAnnotator is depicted in Fig. 2. The area on the left describes the video source and the annotation tracks that are attached to it. On the upper right side there is a list of the current annotations that can be used to modify or delete them. Below there are the video-lenses corresponding to the different annotation tracks. In this example the “Audio visual” video-lens corresponds to a VCR panel to preview the video and the other three lenses, give different perspectives of the video content, namely the script of the movie and the structure in terms of the shot based segmentation. Figure 3 represent a query that was made to the annotated movie. In this example the query combines temporal information (represented by segments with

start and end points in the video stream) and text search in a text based track.

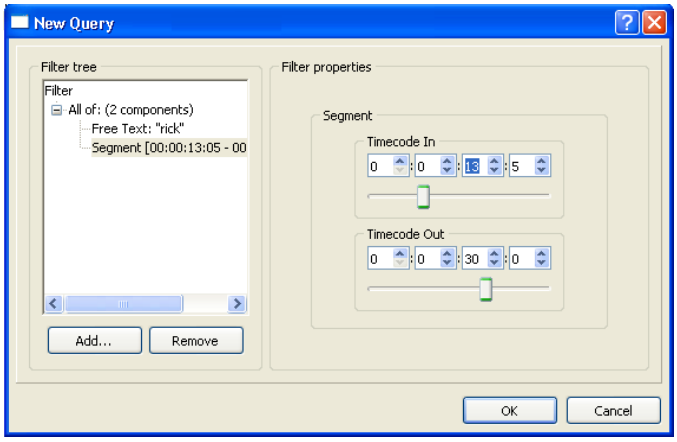

Figure 3: Example of a query

6. CONCLUSIONS AND FUTURE WORK

The VAnnotator was designed in a way that integrates multiple interfaces to the video content. Preliminary tests with users in the project (both in the consumer and professional markets) show that it is possible to accommodate these two types of users. The usage of the MPEG-7 standard as the storage/retrieval format enables interoperability with other annotation systems. Future work on the project includes large scale user testing in order to validate the features that are being currently implemented. Among these features are the collaborative support, relying on a server that will contain video materials and the annotations. The mechanisms for relating several events (or annotations) in the timeline will be used to establish relations between annotation sets belonging to different users. The other main aspect that will be considered has to do with the possibility of each user defining its own video-lenses. There are several ways to obtain the views corresponding to each of the lenses: (1) views can be obtained automatically using cut detection, object tracking or other processing techniques; (2) the user can input the data that describes the video; (3) the data can be imported from an external source, for example, to integrate previously existing annotations made with another system. The component for integrating new lenses, LensFactory, is under development and it will be the basis for a truly adaptive interface, where each user can configure and define the most appropriate views over the video data.

7. ACKNOWLEDGMENTS

This work is partially funded under the 5th Framework program of the European Union within Key Action III (project VIZARD, IST-2000-26354). The VIZARD project is carried out by the following partners: Joanneum Research (A), Technical University of Berlin (D), 4VDO (P), FH Joanneum (A), Forum des Images (F), ORF (A), Duvideo (P), Sony Europe (B).

8. REFERENCES

[1] Nack., F. and Putz, W., Designing Annotation Before It's Needed, ACM Multimedia 2001, Ottawa, Canada, (2001)

[2] Rehatschek, H. and Kienast, G., VIZARD – An Innovative Tool for Video Navigation, Retrieval and Editing, Proceedings of the 23rd Workshop of PVA 'Multimedia and Middleware', Vienna, (2001)

[3] Harrison, B. and Baecker, R., Designing Video Annotation and Analysis Systems, Proceedings Graphics Interface '92, (1992), 157-166

[4] Varlamis, I., Vaziargiannis M., Poulos, P., Akrivas, G. and Spiros, I., X-Database. A Middleware for Collaborative Video Annotation, Storage and Retrieval, in the proceedings of the 8th Panhellenic Conference. Cyprus, (2001)

[5] Davis, M., MediaStreams: An Iconic Visual Language for Video Representation, Readings in Human Computer Interaction: Toward the Year 2000, Morgan Kaufman Publishers Inc., (1995), 854-866

[6] Davis, M., Media Streams: Representing Video for Retrieval and Repurposing, PhD Thesis, MIT Media Laboratory, (1995)

[7] Correia, N. and Chambel,T., Active Video Watching Using Annotation, ACM Multimedia'99, Orlando, Florida, USA, (1999)

[8] Sgouropoulou, C., Koutoumanos, A., Goodyear, P., and Skordalakis, E., WebOrama: A Web Based System for Ordered Asynchronous Multimedia Annotations, WebNet 98 - World Conference on the WWW, Internet, & Intranet, AACE Conferences, Orlando, Florida, USA, (1998)

[9] Elmagarmid, A., Jiang, H., Helal, A., Jishi, A., and Ahmed, M., Video Database Systems: Issues, Products and Applications, Kluwer Academic Publishers, (1997)

[10] ISO MPEG-7. Text of ISO/IEC FCD 15938-2 Information Technology – Multimedia Content Description Interface – Part 2 Description Definition Language, ISO/IEC JTC 1/SC 29/WG 11 N4002, (2001)

[11] ISO MPEG-7. Text of ISO/IEC 15938-5/FCD Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes, ISO/IEC JTC 1/SC 29/WG 11 N3966, (2001)

[12] Bier, E., Stone , M., Pier , K., Buxton , W. and T. DeRose, Toolglass and Magic Lenses: The See-Through Interface, Procdeedings of Sigraph 93, Anaheim, (1993), 73-80

[13] Bargeron, D., Gupta, A., Grudin, J. and Sanocki, E., Annotations for Streaming Video on the Web: System Design and Usage Studies, http://www.research.microsoft.com