Workflow for Integrated Object Detection in Collaborative Video Annotation Environments

Lars Grunewaldt, Kim Möller, and Karsten Morisse

University of Applied Sciences Osnabrück, D-49076 Osnabrück, Germany
http://www.diva.fh-osnabrueck.de

Abstract. Annotation tools for media are getting more and more important. The application for these kind of applications are very manifold. This paper describes some ongoing work on a video annotation environment supporting the media production and post-production process. Based on a former approach, some new development like integrated audio conferencing with recording facilities for audio based work instructions and an automatized video segmentation module are presented. Beside these technical improvements an object based approach for the video annotation process is presented.

1 Introduction

Annotation tools for media, e.g. audio, video or animation are getting more and more attractive and important. Existing approaches and tools are very manifold. There are tools for a collaborative learning approach using hyper-video functionality [3]. Other tools are extending the video information with metadata to make it searchable in a media database, e.g. IBM MPEG-7 Annotation Tool [12]. Moreover, annotation tools can be a useful support in the production process of media. An approach for this kind of annotation has been shown with DiVA [9]. In [7] some important requirements and an approach to use annotations in the production process has been presented. However, during the usage of DiVA some things attracted negative attention. This paper describes some technical improvements in the existing software and shows a way to the next step in video annotation that is an approach to object-based video annotation. All of the annotation tools known so far are using an manual-geometric approach, i.e. annotation in the video plane are defined manually with elementary geometric objects, e.g. circles, rectangles or polygons. It would be very desirable to have an object-based approach, where one can activate a single object in the video layer, e.g. a person or a table. Figure 1 shows the difference between these two annotation approaches.

The organization of this article is as follows. In section 2 the annotation approach of DiVA in video production is briefly described and an overview about some technical improvements is given. In sections 3 and 4 known technology for object-detection from 2D-picture is considered and an approach for object based annotations is presented. Finally in section 5 some concluding remarks and some prospectives for future work are given.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, pp. 565–572, 2006.

(a) Rectangle as a manual-geometric annotation

(b) Object contour highlighting for object-based annotation

Fig. 1. Geometric vs. object-based annotation

2 DiVA – Video Annotation in Post-production

DiVA[1] [9] is a collaboration tool which supports some basic features for synchronized video annotation. Some general requirements for such a tool (see [7]) are: synchronized navigation for several users, definition of textual and graphical annotations which can be archived for work in the subsequent production phases and usage of high-quality video material.

Annotations in DiVA can be defined manually as a text or a graphic object directly on the video layer (manual-geometric approach from figure 1). By a client-server approach several users can navigate in a synchronized manner through the video sequence to define categorized annotations, which are in fact graphical objects combined with textual metadata (see figure 2).

2.1 Technical Improvements for Video Annotation

During the usage of DiVA some things attracted negative attention. These were: manual definition of shots within the video sequence, proprietary format for annotation storage and the lack of an integrated audio conferencing component.

Integrated Audio-Conferencing. As a collaboration tool DiVA has to provide ways of communication to the users while annotating the video content. In a first prototype of the system, such an audio-visual communication was only possible using third party products like external VoIP software or a standard phone line. As this creates a cumbersome overhead of providing a seperate tool next to DiVA itself, integrating a VoIP component is an important feature for synchronized collaboration tools as well for asynchronous tools, if audio recording is considered as well.

[1] DiVA - **Di**stributed **V**ideo **A**nnotation.

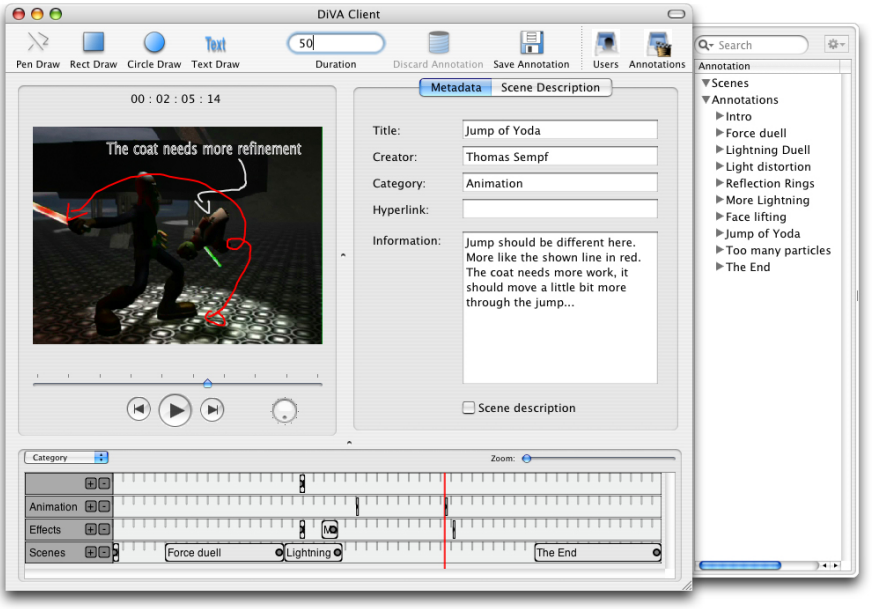

Fig. 2. DiVA - Scene annotation with graphic and text objects, synchronized navigation and annotation archiving

Therefore in [4] a Voice-over-IP component has been integrated into DiVA. Right now, several types of VoIP communication systems and protocols are in wide-spread use; mainly the H.323 protocol suite and the SIP protocol can be found in common VoIP software applications such as Netmeeting or hardware-implemented VoIP components like the Sony PCS series. Some applications are based either on H.323 or SIP, some support both protocols. Additionally, there are several products that are based on propriatary protocols, e.g. Skype. When implementing the audio conference plug-in for DiVA, the ability to easily replace the audio conference unit with another one was very important. Development of VoIP technology is still swift, and it seemed useful to be able to switch between different audio conference implementations to rapidly adopt current communication methods. The current conference plug-in supports the H.323 protocol using a H.323 conference server to manage conferences on the DiVA server side. To match bandwidth requirements different audio codecs are available, e.g. GSM-codecs for lower bandwiths, and G.711 for higher bandwidths. The DiVA client automatically connects to the audio conference system when joining a DiVA annotation session. Once set up, no additional user interaction is needed to activate audio conferencing when connecting to the DiVA system.

For a tool supporting the production workflow management, it is important to track the discussion process about the video content. Therefore it is desirable to provide a functionality to record audio comments of the users. This has been realized by extending the conceptual video annotation process to sessions and

meetings. A conference session can consist of several meetings of different users. For each meeting not only the audio communication is recorded and can be replayed later but also the actual position of video playback during the discussion is stored. This, combined with the possibility of adding timed notes to the conference recording, enables users to use the recordings for later checks of decisions made during a conference or as a basic implementation of a audio-based to-do list or work instructions.

Automated Video Segmentation. An important issue in video annotation is to define annotations for single shots. Shots in a video sequence had to be defined manually in DiVA so far.. In [6] several approaches for an automation in video segmentation were investigated. As a result two modules were developed to split videos into shots: EDL[2]-based segmentation and Automatic Scene Recognition.

The EDL-Segmentation- module loads a EDL-file that can be exported from almost all video-edit tools. An interpreter for the CMX340 and CMX3600 format is implemented. Interpreter for other EDL-formats or more complex file formats like MXF or AAF can be added to the module by corresponding interpreter plug-ins. The interpreter plug-in converts the information found in an EDL-file and creates an annotation for each single shot.

A second and much more flexible module provides a shot segmentation without additional information like an EDL-file. It analyzes the video content for cuts and can be used on all video formats which are compatible to the Apple Quicktime Framework. The price for this very flexible approach in content segmentation is the obligation to analyze each single frame of the video sequence[3]. Off course, for special video codecs dedicated algorithms can be used[4]. They can be added to the segmentation module by a flexible plug-in approach. In [6] several known and some new algorithms for shot segmentation had been investigated. Some of them were integrated in the DiVA-system. With the Contrast Compare algorithm (see [5]) it is even possible to find fades and dissolves. In general, the histogram systems returned equal or even better results than more complex methods like color coherence vectors [11] or principal coordinate systems [10]. A newly developed algorithm is the 2 x 2 histogram, which creates four independent histograms of the four quarters of a frame and compares them with the next frame. The high retrieval rate of shots is a major advantage of this algorithm. However, like in other systems too, rapid light changes can still lead to over-found cuts. But in an application like DiVA this is not a real shortage. The collaborative video annotation is a highly interactive process where the

[2] EDL - Edit Decision List.

[3] This is done by rendering each single frame in a bitmap image and to analyze all pixel information. These pixel information, e.g. histogram, is then compared to the previous frame. In case of a large difference a new shot or segment has been found. The ratio of deviation can be used as a parameter of the segmentation algorithm.

[4] For example codecs based on interframe coding like the codecs of the MPEG-family with techniques like motion-compensation and predictive coding. Here one can also consider the GOP-structure of the video sequence to design much more powerful algorithms for this special case.

users define new annotations or delete existing annotations and shots found by the segmentation module are just an indicator for the collaboration process and can be changed easily.

Standardized Annotation. In the prototype of the software, the annotations were saved in a proprietary XML-format. This circumstance causes a lack of integration and exchange with other software tools. Hence, the MPEG-7 standard had been evaluated for the description of video annotations. This part is still under development.

3 Related Work for Annotation and Object Detection and Recognition

3.1 Annotation Tools

Most of the available systems for video annotation are systems for single user usage and not intended for collaboration with several persons involved concurrently. So far video annotation systems are not known to assist the video production process. The known systems are tools to describe the actual plot of a video. Most of these systems offer a bundle of tools for navigating through the video sequence and adding metadata to it. This metadata is then stored on a server for later retrieval. In some systems MPEG-7 [16] is used for the metadata, e.g. in the IBM MPEG-7 Annotation Tool [12]. This metadata is then saved and could be viewed with a MPEG-7 conformable player.

[3] describes a tool for hyper-video functionality. This approach as well as the Hyperfilm project [13] are extending the hypertext concept to continuous media by defining mouse sensitive regions for one or several video frames which are linked to other information (e.g. pictures, text or other videos). Some other systems for hyper-video are described in [2]. Although hyper-text functionality is an interesting feature, it is not the main focus of the annotation approach considered here.

An interesting project is the filmEd [8] research project of the University of Queensland, Australia. Within filmEd tools for annotating, discussing, and resuming videos are developed. Moreover, software tools to discuss and annotate video sequences in real-time over the Internet in a distributed manner are considered. A prototype called Vannotea has been developed, which enables the collaborative indexing, annotation and discussion of audiovisual content over high bandwidth networks. However, the navigation system is master/client based, that means that only one person can navigate through the video sequence and all clients will be automatically synchronized.

3.2 Object Recognition and Interaction in 2D Pictures

In [1] an overview is given about object detection and recognition in 2D pictures. The presented systems work with a sequence of model points of an object, a degree of invariance (e.g. through rotation or scale) and a description of the

image data, that is e.g. independent to photometrically invariant. Some standard objects like characters or faces can be build in an object database, more complex objects have to be determined by the user in an initial sequence, showing what and where an object is. By-and-by the size of the object database increases and the software, often used in neural networks, can work independent and detect or recognize objects.

But, the enlarging database leads to a slower detection. In movies it is of course possible to use a faster working object tracker to find the object in the next frame again. Nevertheless, to work with objects, the user has to catalog them in a time consuming process. Also it takes time to get a good working system. But even then there will be still errors caused by e.g. multiple objects or partial hidden objects in a frame.

A very interesting approach for object based video annotation will be made possible by a full implementation of the MPEG-4 standard. It offers a lot of additional features for interactive video applications. For instance the user even has the option to move or turn objects directly in a movie. But at the moment these features are usually not implemented. Only a few implementations of MPEG-4 beyond simple audio and video playback are known. Most of them are not available to the public. One interesting approach is the GPAC project [15], that supports at least some of these features in a MPEG-4 production tool and a player. Certainly these features could be used effectively in a collaborative video discussion tool like DiVA.

4 An Approach for Object Based Video Annotation

Recognizing objects in 2D pictures is time consuming and, without any additional information also inaccurate and erroneous. Therefore a new approach is considered in DiVA, because it is a tool supporting the production process. The basic idea is thus not to find objects in existing movies or TV contents, but to get as much as possible information about objects while creating a video sequence or animation with 3D-, videoedit- and/or posteffect-tools.

With a 3D modeling tool like Cinema 4D or a procedural approach (e.g. scenegraph-based systems like OpenSG or Java), one creates a virtual 3D scene but also makes shots of this scene from different camera angles and camera flights in between. Of course, when rendering these different shots, one has all information available for object detection later on. There are two possibilities to store the needed information: coordinates and contours of objects in a rendered frame can be stored as metadata in a MPEG-7 file or objects can be stored directly, when using the MPEG-4 BIFS[5] format. Using the stored data in a video annotation tool, an object contour highlighting, as shown in figure 1(b), can be realized.

In real video sequences with natural actors and objects, only the case when working with more than one video-layer and perhaps the use of chroma keys in editing tools is considered. The contents of the different layers can also be stored

[5] **BI**nary **F**ormat for **S**cenes.

as objects to MPEG-7 or MPEG-4. Certainly, when one video-layer contains more than one object, this method does not guarantee the identification of every single object in that layer.

When rendering frames firstly, then arrange and mix them afterwards with natural video in an editing tool and finally adding some posteffects, the object information, collected during the rendering process, can become invalid if stored to a MPEG-7 file. Thus naturally, the editing- and posteffect-tool also need extension, ensuring the object positions and contours stay valid during the workflow.

5 Summary and Further Work

DiVA is a collaboration platform for video production that supports different steps for a distributed video production process. The distributed approach for navigation, annotation and collaboration on video content makes DiVA unique among other systems.

This paper presents some important improvements like integrated audio conferencing and automation in video segmentation. Especially the later one is a must have feature for these kind of tools. By implementation of a new shot detection algorithm based on partial RGB-histogram values, good automatized recognition results for shots could be reached. The audio-based work instructions are an innovative approach in the workflow of media production. This has been realized by the integration of a audio conferencing and storage component.

Moreover a new approach for object based video annotation has been presented. Work on this feature is still under development and its effectiveness in practice must be shown. But it opens the door for a wide range of new applications. An interesting application of annotated video sequences by an open standard like MPEG-7 is in broadcasting. Regarding DiVA in the production workflow of videos, the created MPEG-7 file can be used for example in interactive MHP applications, when the video is broadcasted. A similar approach is also considered in the GMF4ITV-project [14].

References

1. Amit, Y.: 2D Object Detection and Recognition MIT Press, Cambridge, USA (2002)
2. Dommasch, C.: Entwurf und Realisierung einer komponentenbasierten Anwendung für den kooperativen Wissenserwerb auf Basis von interaktiven Videoinhalten. Diplomarbeit, Fachhochschule Osnabrück, 2005.
3. Finke, M., Balfanz, D.: A reference architecture supporting hypervideo content for ITV and the internet domain. Computers & Graphics 28 (2004).
4. Grunewaldt, L.: Konzeption und Realisierung einer Audio-Konferenz- und Aufzeichnungskomponente im Rahmen eines Video-Annotationswerkzeugs Diplomarbeit, Fachhochschule Osnabrück (2005)
5. Lienhart, R.: Verfahren zur Inhaltsanalyse, zur Indizierung und zum Vergleich von digitalen Videosequenzen, Shaker Verlag, Aachen (1998)

6. Möller, K.: Softwareentwicklung zur automatischen Segmentierung von digitalem Videomaterial, Diplomarbeit, Fachhochschule Osnabrück (2005)
7. Morisse, K., Sempf, T.: An Approach for Video Annotation in Post-Production. In Proceedings of Int. Conference on Computational Science ICCS 2005, Atlanta, GA, USA, May 2005, LNCS 3516, Springer Verlag, 2005
8. Schroeter, R., Hunter, J., Kosovic, D.: FilmEd - Collaborative Video Indexing, Annotation and Discussion Tools Over Broadband Networks. International Conference on Multi-Media Modeling, Brisbane, Australia, January 2004
9. Sempf, T., Morisse, K.: Video Annotation in der Postproduktion. Digital Production. **1** (2005) 103-105
10. Yilmaz, A., Ali Shah, M.: Shot Detection using Principal Coordinate System. IASTED Intl Conference, Internet and Multimedia Systems and Applications Las Vegas, USA (2000)
11. Zabih, R., Pass, G., Miller, J.: Comparing Images Using Color Coherence Vectors Proceedings of the Forth ACM International Conference on Multimedia 96, ACM Press, p. 65-73 Boston, USA (Nov 1996)
12. IBM MPEG-7 Annotation Tool: http://www.alphaworks.ibm.com/tech/videoannex
13. Hyperfilm - Extending hypertext in the video realm: http://www.hyperfilm.it/eng/index.html
14. Generic Media Framework for Interactive Television: http://www.gmf4itv.org
15. GPAC Project: http://gpac.sourceforge.net
16. MPEG (Moving Picture Experts Group): http://www.chiariglione.org/mpeg/