VIZARD[1] - an innovative Tool for Video Navigation, Retrieval, Annotation and Editing

Herwig Rehatschek, Gert Kienast

E-mail: {Herwig.Rehatschek, Gert.Kienast}@joanneum.at WWW: http://iis.joanneum.at
Institute of Information Systems & Information Management,
JOANNEUM RESEARCH, Steyrergasse 17, A-8010 Graz, Austria

Abstract: Non-linear video editing was for several years a domain for professionals due to high prices of capture devices and cameras and difficult to use video manipulation software. With the broad availability of cheap video capture and recording devices video editing became popular also in the consumer market. However, software for manipulating digital video is still complicated to use for home users and non-professionals. We discuss a prototype system for manipulating digital videos featuring navigation, annotation, browsing, search & retrieval. The prototype system was developed throughout an EC project and is currently expanded by an innovative editing concept (video book paradigm) and additional navigation and annotation capabilities (video lens paradigm).

1 Introduction

For several years non-linear video editing was a domain only accessible to professionals users due to difficult use of video manipulation software and large cost of capturing devices and cameras. Now cheap video capture devices and digital cameras made recording and manipulation of digital videos available also for the consumer market. However, even if hardware became affordable for home users, software for manipulating digital video is still complicated to use for non-professionals. Furthermore the functionality offered by commercial tools mainly concentrates on editing of digital videos. Additional functionalities such as navigation, annotation, browsing, search & retrieval needed for a comfortable video manipulation platform are neglected. According to the authors' knowledge there do not exist any systems offering this complete portfolio of features in a user friendly way necessary in order to comfortably manipulate digital videos.

The VideoNavigator developed throughout the VICAR EC project[2] is a prototype system for offering such a portfolio of features, which currently are: navigation, annotation and search & retrieval by identity and annotation search. During the ongoing EC project VIZARD the functionality of the VideoNavigator is significantly expanded and integrated into a user friendly manipulation platform for digital videos suitable for home users. VIZARD will be an open, expandable platform offering modules for editing, annotation, navigation, browsing and retrieval of videos. Furthermore it will provide a suite of plug-ins for automatic extraction of semantic content from digital video streams. The current navigation module allows video content to be organized in a temporal manner or hierarchically according to semantic criteria. A storyboard of relevant keyframes gives a compact overview on the entire video.

In the following we will give an overview of related work in chapter 2. In chapter 3 we will discuss the current existing modules of the VideoNavigator. In chapter 4 the new approach - which will be implemented in the currently ongoing VIZARD EC project - for navigation, editing and annotation videos will be presented. Chapter 5 will give some conclusions and in chapter 6 an outlook for future work can be found.

[1] VIZARD EC project IST-2000-26354
[2] VICAR EC project ESPRIT 24916

2 Related Work

Related work is split into three different domains: video annotation models, content-based image indexing and retrieval and video navigation. All domains are covered by the VIZARD project.

2.1 Video annotation

With the availability of standardized digital video formats and codecs broadcasters are setting up the first huge digital video archives, hence several efforts are undertaken in order to define appropriate data models for storing the associated metadata. Most models cover either low-level (physical) or high-level (semantic) aspects of digital videos [5]. One model for storing a physical, time based representation of digital video and audio was introduced by [7]. General concepts for the physical modeling of digital video and audio data are discussed and a specific model for storing QuickTime movies is introduced. The application of the general concepts allows the specific physical modeling of any other video format. The Layered Multimedia Data Model (LMDM) developed by [8] emphasizes the sharing of data components by dividing the process of multimedia application development into smaller pieces. LMDM claims for the separation of data, manipulation and presentation. Both modeling approaches do not concentrate on the topic of generic film annotation using user definable attributes and values which can be attached to any physical or logical unit (e.g. an act, scene, shot) of a film. Such an approach was proposed by one of the authors in [6] and implemented in a prototype system. The proposed generic annotation model for digital movies allows to structure the film in as many hierarchical levels as needed and to annotate any physical or logical part of the film with generic definable attributes. VIDEX, an integrated generic video indexing model based on low-level visual features and high-level semantic objects was introduced by [5]. VIDEX allows objects to be related to each other spatially, temporarily and spatio-temporarily. The upcoming MPEG-7 standard - where the first Draft International Standard is scheduled for July 2001 - will define a standardized set of descriptors of visual data, a way of defining new descriptors and data structures for indexing and searching by content [17] [18]. The availability of an international standard will be crucial for the development of open systems providing interfaces to existing legacy databases.

2.2 Content-based image indexing and retrieval

Most content-based indexing and retrieval systems work on a query by image example basis. To do so most systems use a shot detection algorithm which also generates one or several representative keyframes [2]. These keyframes are then indexed as still images, often motion information is added on a per-reference frame basis. Examples of such systems are QBIC [9], VisualSeek [10] and Virage [11]. In the meantime the Virage system has evolved to a commercial company selling products in the video indexing domain [12]. Their media analysis software supports face recognition [13], text recognition [14] speech-to-text transcription, extraction of teletext and closed caption information. In 1994 the Motion Content Analysis (MoCA) project under the guidance of Prof. Dr. W. Effelsberg was started at the University of Mannheim in Germany. During the past years, different applications have been implemented and the scope of the project has concentrated on the analysis of movie material such as can be found on TV, in cinemas and in video-on-demand databases. Analysis features developed include four domains: (1) features of single pictures (frames) like brightness, colors, text, (2) features of frame sequences like motion, video cuts, (3) features of the audiotrack like audio cuts, loudness and (4) combination of features of the three classes to extract e.g. scenes [1], [2], [3], [4], [5].

2.3 Video navigation

A system which allows a navigation within video essence is the Hyper-G system [15] which was initially developed at the Graz University of Technology and is now commercially sold as Hyperwave [16]. The Hyper-G system supports the definition of hyperlinks as known from the

World Wide Web within digital videos in order to reference to other information stored on the server, i.e. additional text information on the current scene etc.

3 Video Annotation, Navigation and Retrieval

The overall goal of the above-mentioned VICAR project (Video Indexing Annotation And Retrieval) was to build a framework that allows to easily structure, annotate, store and retrieve video content. The developed system, the VideoNavigator (VIN), was designed as a prototype of a media asset management system that would be used in a television archive to manage the entire archive holdings.

The main motivation for the VICAR project was an increasing demand for storage and content management, (semi-)automatic annotation and search and retrieval facilities. This demand has arisen from the process of 'traditional' television archives turning into digital multi-media archives.

During development of the VideoNavigator the emphasis was on creating open interfaces in order to allow extending its functionality very easily. Therefore a plug-in architecture and an open XML-based file format were used.

3.1 Annotation

The VideoNavigator provides the functionality to automatically extract meta data from digital video content. This meta data extraction ranges from structuring the video into sequences of the video to creating textual annotations. This is done by so-called analysis plug-ins, each of which is responsible for creating a certain class of annotations. So far there are the following analysis plug-ins available:

- *Basic image analysis:* Performs detection of shot boundaries (cut detection), keyframe extraction and creates the stripe image. The cut detection is based on calculation differences between consecutive frames using dynamic thresholds to improve results
- *Motion analysis:* Extracts motion information (e.g. camera zoom, pan and object motion) Motion information is determined directly from the motion vectors available in the MPEG file
- *Identity matcher:* Performs visual feature extraction (color distribution, image structure) which allows to search for visually similar images at a later point in time. This plug-in creates binary information and thus its results are not visible in the user interface
- *Face finder:* Detects occurrences of faces in a video
- *Setting classifier:* Determines the setting (inside, outside, forest etc.) of a scene

The VideoNavigator allows to view all annotation classes separately or combined. The first option allows for e.g. view a shot list, or a list all of faces occurring in the video. The second method merges the shot list with all annotations and hence gives an overview of all annotation related to a shot (see Figure 1).

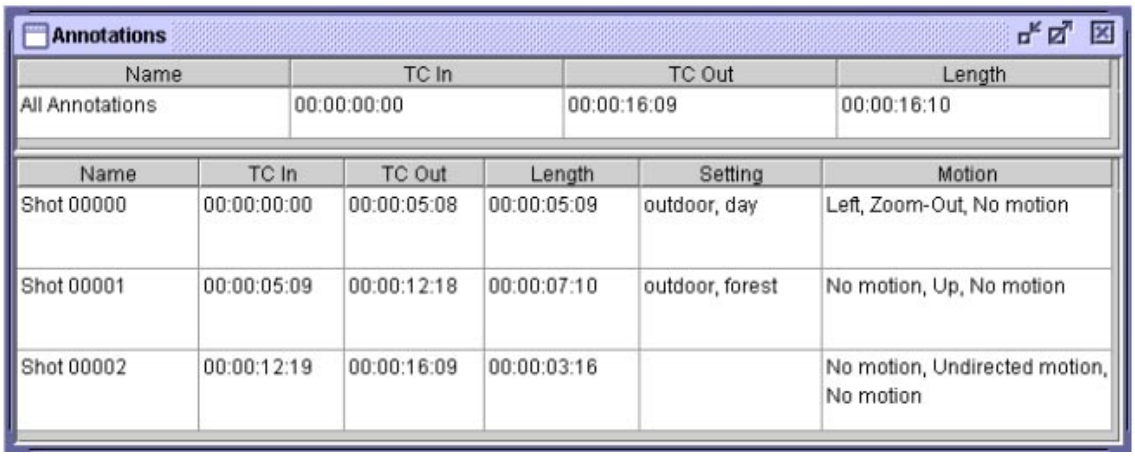

Figure 1: combined visualization of the annotations

3.2 Navigation

Besides a simple video player the VideoNavigator provides four different navigation aids to quickly navigate through video content:

- The *structure tree* allows to hierarchically structure the video in either a time-related manner or due to semantic context (e.g. combine all sequences about a certain topic into one group).
- The *annotation table* provides a list of all sequences in a node selected in the structure tree. Besides the name of the sequence it provides in- and out-timecodes of the sequence, length of the sequence and all related textual annotations if available. It also allows to modify existing annotations or creating new annotation classes.
- The *keyframe panel* displays the “storyboard” of the video. Keyframes are extracted by the basic video analyser plug-in. These keyframes give a compressed overview about the content of the video.
- The *stripe image* is an even more compressed representation of the original video than the keyframes are. It is created by adjoining the middle vertical column of pixels of every video frame. This produces an image where the x-axis represents time dimension. Effective use of this stripe image requires a little training but in many cases it gives very quick information about video activity. A typical stripe image is given in Figure 3.

All four of this navigation aids are fully synchronized, i.e. if the user clicks on a keyframe the corresponding shot is highlighted and the stripe image and the video player are set to that position as shown in Figure 2.

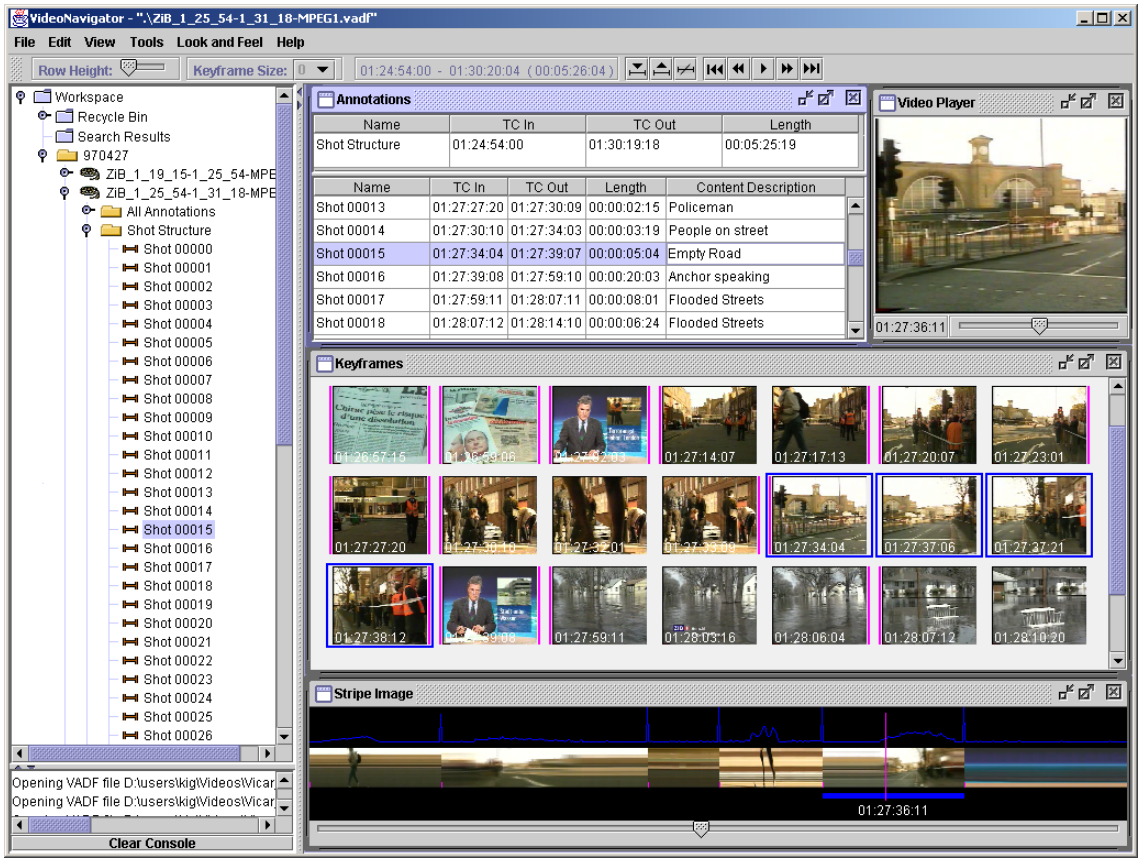

Figure 2: VICAR VideoNavigator and its main components: structure tree (left), annotation table (top center), video player (top right), keyframe panel (center) and stripe image (bottom).

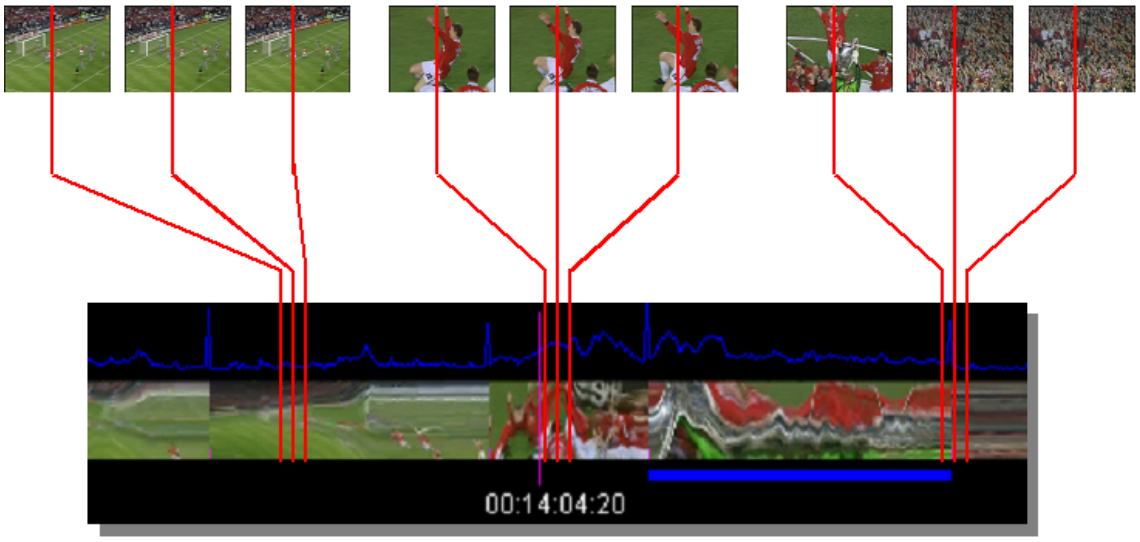

Figure 3: stripe image

3.3 Retrieval and Annotation

The VideoNavigator allows retrieval of video content by two different search methods:

- Search within textual annotations. This is supported by a thesaurus, which supports synonyms (e.g. if the user searches for ‘car’ the thesaurus will expand the search also to the terms ‘truck’ and ‘automobile’).
- The identity matcher module allows the user to specify an image (either a keyframe from the database or any image from a file) and the system retrieves all visually similar images stored in the database.

Technology developed during the VICAR EC project is currently exploited together with a two commercial companies. More details are given in chapter 6.

4 A new Approach for Navigation, Editing and Annotation

Within the VIZARD project we will develop a new generation video publishing tool. The main objective of VIZARD is to introduce a new user friendly concept to deal with video content. A video editor should not limit the creativity of users to simple time-line representations, but they should have the possibility to manipulate the inner structure of a video document (like it can be done in an electronic text document) and work with hyper-video representations (see below) to build distributed and personalized collections. Publishing personalized video content should be as easy as writing a text document.

The main components of VIZARD are:

- video navigation by providing an intuitive desktop to navigate through video documents, preview video footage and development of compact representations and a temporal segmentation of a video which will allow to explore the content and structure of a video in a very fast way.
- video editing functionality based on the video book paradigm including planning and drafting (pre-production), content selection (structuring & reviewing) and arrangement (comparable to traditional editing).
- a video annotation wizard which is based on the video lens paradigm (annotation format compatible to MPEG-7).
- a core library supporting video editing in the compressed domain by providing "faster than real time" processing of video content

The main components of the VIZARD software architecture are presented in Figure 4.

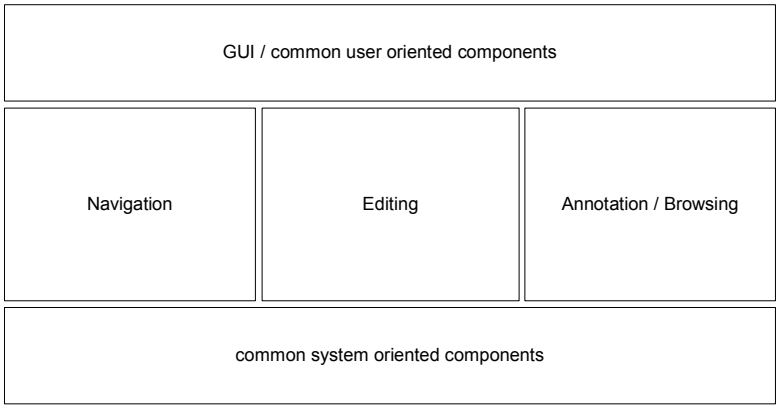

Figure 4: main modules of VIZARD

The three main components are the navigation, editing and annotation / browsing. All of these components are independent from each other as far as possible and thus only communicate via well defined APIs or files. All components appear to the user in a coherent design. There are also common components which are divided into system oriented and user oriented parts. The possible interaction between components are specified in well defined interfaces. Each of the three main components provides an API which may be used from the other components, but will also work alone. Next to the API communication the three main components are sharing their (meta) data. So for example the navigation part can use the annotation data by accessing the annotation metadata. The (meta) data are well defined and commonly accessible. A dataflow diagram of the VIZARD components is given in Figure 5.

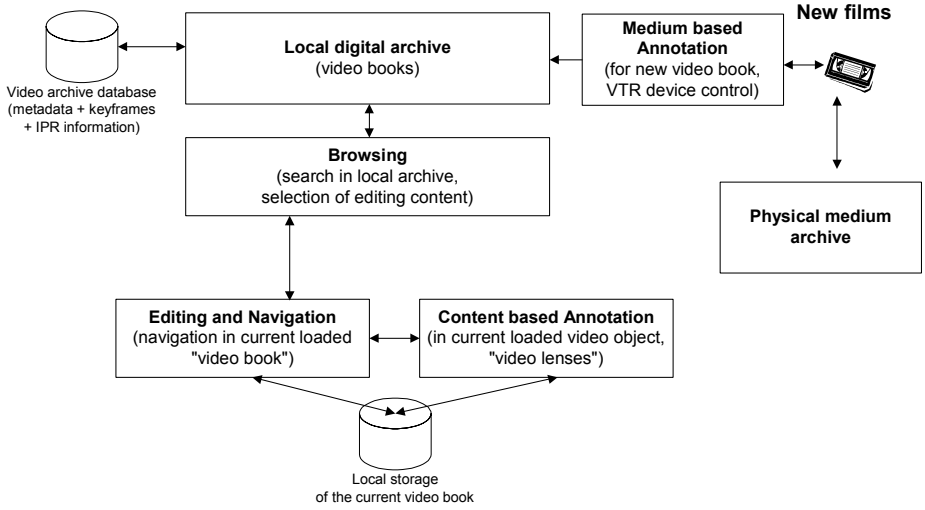

Figure 5: dataflow between the main modules

4.1 Navigation

The main objective of the navigation module of VIZARD is to provide the user with a comprehensive "Video collection search & management tool". Hence the navigation module will provide functionalities in order to maintain a considerable small, local archive of stored "video books" and functionalities in order to comfortably navigate within one and between several video books. Hence the navigation component will cover the "local digital archive", the "browsing" component and partially the navigation within the editing component, given in Figure 5. The local archiving component shall organize content by hierarchically structure-able collections containing video books as objects. We plan to realize the following features within the navigation module of VIZARD:

- *Import* (e.g. from DV cameras,) and *export* (MPEG-7 / XML metadata and video essence) modules.
- *Pre-processing* (import of raw video including shot detection, key frame and stripe image generation, camera motion extraction and quality check) resulting in an empty video book.
- The possibility to *organize the collections* (i.e. by drag & drop) according to user needs (this feature is needed by journalists in order to pre-process search results and make up a story).
- A *search facility* within the video book (and local archive or collection) featuring: text search (annotations - connection to annotation toolkit), identity search, and search for video sequences. Search results are stored in a new collection containing video books.
- *Annotations* on the video book.

- *Structuring* of video books - i.e. making chapters, sub-chapters, table of contents, summary, etc.
- Provide *navigation* using this structure (comparable to the table of contents in text documents).
- *Drag & drop* between video books.
- Provide a "*Hyper-video representation*" of videos by letting the user define hyperlinks within one video book to target-able objects. Such objects can be a video book or a location within a video book, annotations, or any additional information such as text documents, images, etc. This feature allows to create simple storyboards.
- Provide *different views* of the book, i.e. zoom in/out, view different number of keyframes, visualization of stripe image with different parameters and so on.

4.2 Editing

The basic idea of the VIZARD video editor module is to move away from a simple time line representation as it is currently used in commercially available video editing systems. We want to provide an overall visualization of the object structure and allow for inner manipulation of the video document [25] by generic layout elements (styles). Hence the basic idea of the VIZARD video editor is to organize videos like a book in chapters and sub chapters, add meta- and layout-information to videos and finally definition of styles, (e.g. wedding video style, vacation video style) which can be applied directly to videos (like in a text processor). This idea is outlined in Figure 6 and further on referred to as the "video book paradigm" [23], [24].

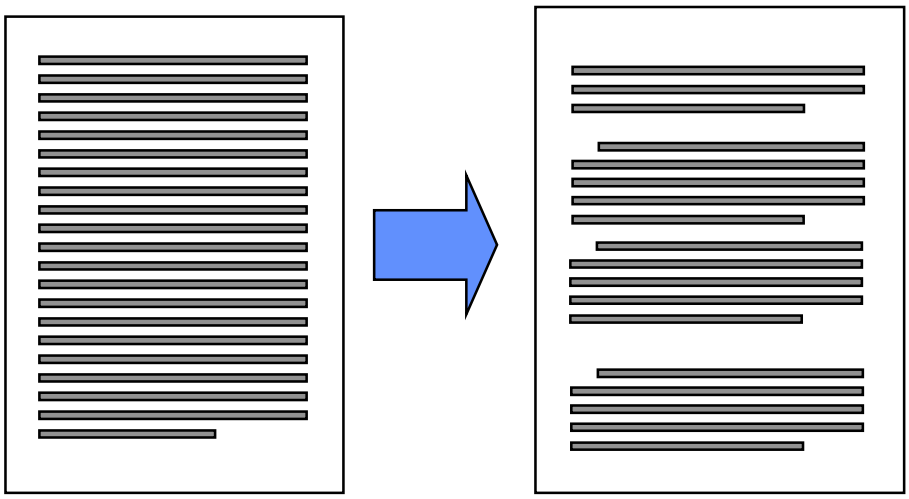

Figure 6: video editing according the "video book" paradigm

For the editing functionality algorithms for *compressed-domain video editing and manipulation*, as for example described in several publications of Meng/Chang [19], [20] will be applied. Very high throughput can be achieved by operations, which are directly applied to the compressed data, and by updating only parts (in the temporal and spatial dimension) of a video sequence where a change has been applied. *Video rendering* will be no more a real-time or near real-time process, but similar in terms of speed to the generation of a "ready to print" text document. This can be compared by generating a Postscript file from which a print document can be generated. In connection with the video book paradigm this means we first will generate a pre-format which still contains all additional information from the video book and can be read by other systems as well. This format can then be directly converted into the video stream which was intended to be generated by the author (rendered video).

4.3 Annotation and Browsing

Most current approaches to video annotation are either based on general and flat textual annotations, or propose complex languages for video annotation (for example, "Media Streams" from MIT Media Lab [21]). In fact annotation is a broad concept that ranges from the registry of physical characteristics (like color or special effect) to content related aspects (like outdoor scene, person speaking, dog sleeping). The VIZARD annotation module will offer flexible structuring and composition of annotation mechanisms. The design rationale for the annotation wizard functionality is: there is no universal annotation scheme, and many of the possible annotation schemes have to be combined for an effective annotation and later retrieval. Moreover, classification and annotation schemes should easily be applicable without demanding painful learning curves for the users. This will be achieved by the "video lens" mechanism which provides a user individual perspective and adapts seamlessly to the standards required by video storage and retrieval systems.

A video lens represents a group of annotation attributes out of a common set of attributes (i.e. the SMTPE attributes) or classification scheme which was individually put together by a user. Examples of these are: physical video attributes (format, quality, color temperature), content attributes (talking heads, outdoors), styles and/or editing + composition attributes (documentaries, music clips, news programs, etc.), standard schemas e.g. ISAD (International Standards for Archival Description - see [22]), standard annotation and international standards like MPEG-7, and specific in-house and private metadata schemas, targeted for specialized groups and/or activities in the video business value chain. Video lenses can hence be used also for filtering by applying special groups of attributes on existing, complex annotations.

5 Results and Conclusions

In this paper we described a prototype system which allows to analyze, annotate and navigate through digital videos. Video analysis includes shot detection, keyframe extraction, motion analysis, identity search and a face finder. Navigation is supported by four different elements, a hierarchical structure tree, an annotation table, a keyframe panel and the stripe image which are fully synchronized.

The *stripe image* and the keyframes which are both compressed representations of video content turn out to be value navigation aids. Effective use of the stripe image does require some training but it provides useful information about video content and can be used for navigate through a video.

Furthermore we gave some new ideas - currently implemented in an ongoing EC project - for manipulating digital video streams. First the *video book* paradigm was introduced. The basic idea of the video book paradigm is to move away from a simple time line representation towards a visualization of the inner structure of a video similar to a book. The video book will provide an overall visualization of the object structure and allow for inner manipulation of the video document by generic layout elements (styles). Hence videos can be organized like a book in chapters and sub chapters, where meta- and layout-information can be added and finally definition of styles can be done.

The concept of *video lenses* was introduced. The basic idea for video lenses can be summarized as follows: since there is no generally applicable annotation scheme many of the possible schemes have to be combined for effective annotation and retrieval. Furthermore it should be possible to apply annotation and classification schemes intuitively for non-professional users.

This goal will be achieved by using the *video lens* paradigm which provides an individualized perspective and adapts seamlessly to standards required by video storage and retrieval systems. A video lenses represent a groups of annotation attributes or classification schemes which can be individually created by a user. They can hence be used also for filtering by applying special attribute groups to existing, complex annotations.

6 Outlook

The modular system architecture of the VideoNavigator does not only allow to easily extend the system's functionality, it also supports reusability of code. At the moment there is a cooperation with Virage, Inc. a California based software company that provides products and services for digital media asset management. Virage's VideoLogger® [12] is a system for ingesting digital video into a database which also automatically extracts meta data (e.g. speech-to-text transcription, OCR, extraction of teletext and closed caption information, face recognition etc.)

Based on the shot detection algorithm developed in the VICAR project there is now the media analysis plug-in ShotLogger® available allowing the Virage VideoLogger® to automatically structure a video into shots.

Another technology currently being adapted for the Virage framework are the identity indexer and identity matcher modules. The SimilarityLogger® plug-in creates the feature vectors which are scanned by a prototype application which allows to retrieve video content visually similar to given material.

Another application of VICAR technology is used at the archive of the Austrian Broadcasting Corporation (ORF) which uses VICAR's similarity search technology since November 1999 to index their entire news material.

7 Acknowledgments

This work is partially funded under the 5th Framework Programme of the European Union within Key Action III of the IST Programme (project "VIZARD" IST-2000-26354). The VIZARD project is carried out by the following partners:

Technical partners:

- JOANNEUM RESEARCH - Institute of Information Systems & Information Management (A), responsible for the navigation module & project co-ordination
- Technical University Berlin - Prozessrechnerverbund-Zentrale (D), responsible for compressed editing toolkit
- 4VDO – Sistemas e Servicos Multimedia (P), responsible for the annotation module
- FH Joanneum – Studiengang Informations Design (A), responsible for the editing moduleEnd-user partners (definition of user requirements, usability, pilot operation and testing):
- Forum des Images (F)
- Austrian Broadcasting Corporation ORF (A)
- Duvideo II S.A. (P)
- Sony Europe (B)

8 References

[1] R. Lienhart. "Comparison of Automatic Shot Boundary Detection Algorithms", Image and Video Processing VII 1999, Proc. SPIE 3656-29, 1999.

[2] R. Lienhart, W. Effelsberg, R. Jain. "VisualGREP: A systematic method to compare and retrieve video sequences", Technical Report TR-97-005, Praktische Informatik IV, University of Mannheim, October 1997

[3] R. Lienhart, C. Kuhmünch, W. Effelsberg. "On the Detection and Recognition of Television Commercials, Proc. IEEE Conf. on Multimedia Computing and Systems, Ottawa, Canada, pp. 509 - 516, June 1997.

[4] B. Davies, R. Lienhart, B.-L. Yeo. "The Video Document", Multimedia Storage and Archiving Systems IV, SPIE Vol. 3846, pp. 22-34, 20-22 September 1999.

[5] R. Tusch, H. Kosch, L. Böszörmenyi. "VIDEX: An Integrated Generic Indexing Approach", ACM Multimedia Conference 2000, Los Angeles (USA), October-November 2000.
[6] H. Rehatschek, H. Müller: "A Generic Annotation Model for Video Databases", D. P. Huijsmans, A. W. M. Smeulders (eds) Lecture Notes in Computer Science vol 1614 Visual Information and Information Systems Third Int Conf, Visual'99, Amsterdam, June 1999
[7] Ch. Breiteneder, S. Gibbs, D. Tsichritzis. "Modelling of Audio/Video Data", pp. 322-339, Karlsruhe, Germany, 1992
[8] Schloss, Wynblatt. "Providing Definition and Temporal Structure for MM data", Proc. of the Second ACM Int. Conf. on MM, CA., ACM Press, ISBN 0-89791-686-7, S. Francisco, 1994
[9] M. Flickner, H. Sawhney, W. Niblack et.al. "Query by Image and Video Content: the QBIC System", Computer, vol. 28 (no.9), p.23-32, 1995
[10] J.R. Smith, S.F. Chang. "Visual SEEK: A fully automated content based image query system", ACM Multimedia conference, Boston, MA, 1996
[11] A. Hampapur, A. Gupta, B. Horowitz, et.al."Virage Video Engine", Proc. SPIE Vol. 3022, p. 188-198,Storage and retrieval for Image and video databases V, K. Ishwar, R. Jain;Eds, 1997
[12] Virage Ltd. "VideoLogger", <URL: http://www.virage.com>, 2001
[13] Visionics Corporation. "FaceIT", <URL: http://www.visionics.com/>, 2001
[14] SRI (Stanford Research Institute) International. "CONText", <URL: http://www.sri.com/index.html>, 2001
[15] H. Maurer, K. Andrews, F. Kappe. "The Hyper-G Network Information System", J.UCS 1.4, pp. 206-220, 1995.
[16] H. Maurer, J. Lennon. "HyperWave: The Next Generation Web Solution", (Ed.), Addison-Wesley Longman, London, 1996.
[17] MPEG Requirements Group. Applications for MPEG-7. Document ISO/MPEG N2084, MPEG San Jose Meeting, 1998.
[18] Murase, H. and Nayar, S. Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, 14, pp. 5-24, 1995.
[19] J. Meng and S.-F. Chang, Tools for Compressed-Domain Video Indexing and Editing, Proceedings, IS&T/SPIE Symposium on Electronic Imaging: Science and Technology (EI'96) - Storage & Retrieval for Image and Video Databases IV, Vol. 2670, San Jose, CA, February 1996
[20] J. Meng and S.-F. Chang, CVEPS: A Compressed Video Editing and Parsing System, Proceedings, ACM Multimedia 96 Conference, Boston, MA, November 1996
[21] Davis, Marc. "Media Streams: An Iconic Visual Language for Video Representation." In: Readings in Human-Computer Interaction: Toward the Year 2000, ed. Ronald M. Baecker, Jonathan Grudin, William A. S. Buxton, and Saul Greenberg. 854-866. 2nd ed., San Francisco: Morgan Kaufmann Publishers, Inc., 1995.
[22] ISAD, "International Standards for Archival Description" <URL: http://dobc.unipv.it/obc/add/infap/archdes/isad(g)e.html>, 2001
[23] Müller H., Tan E.S.. "Movie Maps", in Banissi, et al. (Eds.) Proceedings of the 1999 IEEE International Conference on Information Visualization - IV99, London, 1999, pp. 348-353.
[24] Müller H., Tan E.S. "Visualizing the Semantic Structure of Film and Video" to appear in Proceedings of the Electronic Imaging 2000/Visual Data Exploration and Analysis.
[25] Bob Davies, Rainer Lienhart, and Boon-Lock Yeo. "The Video Document". Multimedia Storage and Archiving Systems IV, SPIE Vol. 3846, pp. 22-34, 20-22 September 1999; also Technical Report MRL-VIG99014, June 1999.