Herwig Rehatschek, Werner Bailer, Helmut Neuschmied, Sandra Ober and Horst Bischof

A Tool Supporting Annotation and Analysis of Videos

1. Introduction

Since the early 1990s, the Institute of Fundamental Theology of Graz University has investigated the relationship between theology and media, especially film, in the context of its research focus Theology – Aesthetics – Visual Arts – Film – Culture. It is also a member of the International Research Group *Film and Theology*.[1] An important focus of its research has been the detailed analysis of videos with respect to religious symbols. In addition, it explores the relationship between media and society, media and the construction of reality, the rise of the religious in media society, and the body as a site for religious experiences in films. Research in this area is not only performed by Graz University but also on an international level. Several well-known academics and institutions, for instance, are Peter Horsfield, an Australian theologian, Stewart Hoover and the research center at the University of Colorado at Boulder,[2] or the Center for Religion and Media at New York University.

In order to effectively perform research in the area of theology and video (film), efficient video annotation tools are needed.[3] The main genre of interest for institutions like the Institute of Fundamental Theology are feature films. Unfortunately, especially in this genre, practically no tools exist to support the researcher with video annotation tasks. 'Video annotation' in this context refers to adding high level semantic information (usually in text form) to the video stream. The advantage is that annotation makes hidden information explicit and facilitates the analysis of the video. Due to the lack of tools until now, the first research work at the institute has been performed with a simple video recorder, a big pile of paper with annotations to specific parts of the film referenced by time codes and a lot of spooling. This is very time consuming. The main reason for the lack of tools in the feature area is that this genre is of minor interest for

1 Cf. http://www.film-und-theologie.de.

2 S. M. Hoover, K. Lundby, Introduction. Setting the Agenda, in: S. M. Hoover, K. Lundby (eds.), *Rethinking Media, Religion, and Culture*, Thousand Oaks: Sage, 1997, 9.

3 Cf. Christian Wessely's article in this volume.

broadcasters and archives. In comparison to news and sports material, the reuse of parts of feature films is very complicated due to copyright restrictions. In the area of sports and news, the level of reuse is considerable, therefore commercial systems with a high grade of automation already exist in this area.

However, those systems are not really suitable for the detailed analysis of feature films since most tools only support global annotations valid for the entire content, e.g. film title, abstract, actors, etc. For a detailed analysis of a film, however, the researcher has to be able to work with smaller units, such as shots and scenes, be able to annotate them (temporal and spatial annotations) and relate them to each other. Temporal annotations describe a film segment which is defined by a start and end time code. These segments can either be a scene or a shot, but also any arbitrary part of the film (e.g. the appearance of a specific person or object or annotations referring to locations etc.). Spatial annotations describe the content of one frame, and hence the semantic decomposition of the image by describing the persons and objects present in this single frame.

A crucial aspect in the design of an efficient annotation tool is its ability to support the researcher (or 'human annotator') in usually time-consuming tasks by offering as many automatic features as possible. Annotations to be performed fully or at least semi-automatic may include shot detection, camera movement, extraction of relevant key frames, extraction of color similarity features, object detection, object recognition, and transformation of spatial annotations into temporal ones. Spatial annotations can be transformed into temporal ones by object tracking, i.e. by following objects automatically. A person, for example, may be marked by a rectangle in one frame and then automatically be tracked forwards and backwards. This will relieve the annotator from the tedious task of marking one and the same person in hundreds of frames and save valuable time.

In this paper, we introduce our Semantic Video Annotation Tool (SVAT) application which enables efficient annotation combined with a set of automatic annotation plug-ins. All metadata of the tool is stored in the ISO standard MPEG-7[4]. There are usually two main stages in the process of video annotation: structuring of the video into a list and hierarchy of segments, and the addition of annotations of various types to these segments. Both steps can be supported by automatic tools. The complexity of automatic annotation and the need for user intervention is related to the level of the feature to be annotated. Low-level features are those that can be derived directly from the audiovisual data (e.g. "15% of the screen area is black", "the actress's voice has a fundamental frequency of 250Hz"). Thus, their description is unambiguous and can be easily automatized. The higher the level of the feature, the more semantics it conveys

4 ISO/IEC 15938:2001.

and the harder it is for an automatic annotation, as additional knowledge is required. This knowledge can be the prior knowledge of the viewer (knowledge the viewer already possesses, e.g. we recognize a table because of our everyday experience), domain knowledge (expert knowledge in a specific domain, e.g. for a referee in a soccer game, a ball rolling across a line between two goal posts is an important event), or contextual knowledge (knowledge we gather from accompanying actions and the context, e.g. we recognize the robot R2D2 in the end of a given STAR WARS film because it was introduced in early scenes of the film). The need for acquiring, representing, and using this additional knowledge complicates the automation of annotation for high-level features.

The problem of describing semantics is especially evident in terms of content structuring. Segmenting a video into shots is a comparatively simple task, which can be performed satisfactorily with automatic tools. This is due to the fact that the concept 'shot' has a very clear definition. This is no longer true, however, for larger units, such as scenes. A scene is often defined by the unity of time and place;[5] for example, the unit of continuous action at a location or following a character (with exceptions to this continuity in some cases),[6] and is required to be homogenous in style (positioning of actors, visual axis, etc.).[7] This definition is already somewhat fuzzy from a technical point of view. Additionally, we have to be aware that when we speak of place and time, this refers to the place and time perceived by the viewer. There may be subsequent shots that have no objective spatial or temporal relation, but if montage is used to make the viewer believe that the action is taking place at the same location or at the same time, the viewer will perceive the separate shots as being part of one scene. If we see content analysis as reversing the authoring process, shot detection merely reverses the technical act of cutting, while scene detection requires reversing the creative process of montage. Thus, content structuring is one of the tasks that cannot be fully automated, but tools can support the researcher in order to increase the efficiency of the manual process.

In the following, we first review related work. In section three, we describe the main parts of our video annotation tool. Section four presents the results of our work and offers some conclusions, while section five focuses on further developments.

5 Merriam-Webster's *Collegiate Dictionary*, 11th edition, 2003.

6 R. Arnheim, Film als Kunst, in: *Texte zur Theorie des Films*, 3rd ed., Stuttgart: Reclam, 1998, 189f.

7 D. Arijon, *Grammar of the Film Language*, Los Angeles: Silman-James Press, 1991, 20ff.

2. Related Work

There is a large amount of work dealing with the description of audiovisual media, but most of it concentrates on two genres: news and sports broadcasts. One reason for this phenomenon is the commercial relevance, as segments of sports and especially news content are frequently reused after their production (e.g. when the Concorde crashed, hours of ad hoc programs had to be filled with archive documentation material of the Concorde, because only a one minute sequence from the actual crash existed) and initial airing, so that they are valuable assets for broadcasters. Segments from feature films are hardly reused in other contexts, so that a detailed annotation is commercially not interesting. Further, compared to feature films, news and sports broadcasts have very clear dramaturgical structures, which makes the automation of segmentation (for example of news stories) more feasible.[8] There are only a few works dealing with general segmentation in feature films, most of them using rather simple features like dynamics (film tempo),[9] or defining "computable" units of the content[10] that do not necessarily coincide with scenes or sequences. There are, however, many low- and mid-level content analysis tasks that can be successfully automated (cf. Section 3.1), and even approaches for the automatic detection of some types of scenes that have clear visual or audio patterns, such as dialogs.[11]

With the advance in image processing, information retrieval, and database management, content-based image retrieval (CBIR) has been actively studied in recent years, which has resulted in a number of systems and techniques, both in academic and commercial domains. The goal of such a system is to find, given a specific image, the closest matches to that image in the video which itself can be seen as a large set of images, by looking at the image content. Many such systems exist but IBM's Query by Image Content (QBIC)[12] and Virage's VIR engine[13] are probably the best-known commercial systems. The user can specify

8 W. Kraaij and J. Arlandis, TRECVID-2004 Story Segmentation Task: Overview, *Proc. of TRECVID Workshop*, Gaithersburg, Nov. 2004.

9 B. T. Truong, S. Venkatesh and C. Dorai, Film Grammar Based Refinements to Extracting Scenes in Motion Pictures, *IEEE International Conference on Multimedia and Expo*, Lausanne, 2002, 281–284.

10 A. Hanjalic, R. L. Lagendijk and J. Biemond, Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, nr. 4, June 1999, 580–588; H. Sundaram and S.-F. Chang, Computable Scenes and Structures in Films, *IEEE Transactions on Multimedia*, vol. 4, nr. 4, Dec. 2002, 482–491.

11 G. Haberfehlner, *Development of a System for Automatic Dialog Scene Detection*, Diploma Thesis, Fachhochschule Hagenberg, 2004.

12 Flicker M., Sawhney H. et al., Query by image and video content: The QBIC system. *IEEE Computer*, 28(9), 1995, 23–32.

13 Virage, http://www.virage.com.

a number of parameters prior to the searching. CBIR systems are applications often specialized in processing large still-image databases. The techniques applied in this scenario can also be used for image retrieval in videos because a video can be seen as a large set of single images. Most of these CBIR systems rely on the use of low-level image information, such as color,[14] shape,[15] and texture[16] features, to search for images in the video database. Recently, content analysis methods have focused on temporal information using motion features. With the help of the motion content, a hierarchical structure is built for a given database for indexing and retrieving video shots. In general, using motion features for representing video content is useful for some parts of a video, which have certain types of motions. However, within the motion class, there may be many subclasses that cannot be satisfactorily separated by motion features.

Recent studies in CBIR have focused on the approaches based on relevance feedback, which tries to bridge the gap between the high-level concepts and low-level feature representations by modeling the user's subjective perception from the user's feedback.[17] High-level concepts can be learned offline (i.e. preprocessing, before – and not during – the search takes place), and can be utilized and refined based on the user's specific interest during the on-line retrieval process. Relevance-feedback-based systems have two major limitations. These approaches estimate the ideal query parameters only from low-level image features. Due to the limited power of these features in presenting high-level semantics, it may not be effective in modeling user's perceptions. The feedback information provided in each interaction contains high-level concepts which can solely be used to improve the current query results for a specific user.

One well-known object retrieval approach in videos is known as *Video Google* and was popularized by J. Sivic.[18] The task is to search for further similar appearances of the requested object in the entire video. The object is selected by defining a region of interest within a frame of the video. Because the region of

14 R. O. Stehling, M. A. Nascimento, A. X. Falcao, On Shapes of Colors for Content-Based Image Retrieval. In: *ACM International Workshop on Multimedia Information Retrieval* (ACM MIR'00), Los Angeles, 2000, 171–174.

15 D. S. Zhang, G. Lu, Generic Fourier Descriptors for Shape-Based Image Retrieval. In: *Proceedings of the IEEE International Conference on Multimedia and Expo* (ICME'02). Lausanne, Aug. 2002, 425–428.

16 L. M. Kaplan et al., Fast Texture Database Retrieval Using Extended Fractal Features. In: *Proceedings of IS&T/SPIE Conference on Storage and Retrieval for Media Databases*, 1998, 162–173.

17 L. Goldmann, L. Thiele, T. Sikora, Online Image Retrieval System Using Long Term Relevance Feedback, *CIVR 2006*, LNCS 4071, 2006, 422–431.

18 J. Sivic, A. Zisserman, Video Google: A Text Retrieval Approach to Object Matching in Eideos. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2, 2003, 1470–1477.

interest is not limited, the user can search for similar appearances of an arbitrary object. To speed up the search process, the video is analyzed in a thorough and time consuming offline pre-processing task without any user action needed.

The task of content annotation is to enrich the audiovisual content metadata, i.e. data *about* the content, and data that *describes* the content. The metadata can be extracted from the content itself or come from external sources (e.g. scripts). Content analysis can thus be seen as reversing the authoring process,[19] during which an audiovisual material is created based on information about the content to be produced (e.g. script, storyboard, scene sketches). Three perspectives for analyzing audiovisual content have been proposed:[20] layout (e.g. shot structure, camera motion), content (e.g. people and objects appearing) and semantics (e.g. scenes, named events). These three perspectives also correspond to the levels of features that can be extracted from audiovisual content, ranging from low-level features describing layout to high-level features describing semantics. The automatically extracted features can be used to support a human annotator, when an efficient annotation tool is available that visualizes the extracted features. Those manual annotation tools proposed in literature, which are relevant in the context of this work, are discussed in the following section.

The *Deconstructor*[21] is an online film analysis tool which assists the exploration of cinema's visual syntax through its arrangement and organization of shots. The Deconstructor can be used to view and dissect film scenes into a series of shots, allowing one to focus on examining the components of the whole in order to layer and juxtapose variables along the time line. One can view film scenes, extract a series of shots from the scene, and provide a description of each shot by considering the common variables such as shot type, time and angle. This information is then used to visualize the shots using a graphical representation of their descriptions. This aims to provide a sense of score for each scene, material that probes further analysis, and helps the user to identify rhythms and cycles.

Anvil[22] is a publicly available research tool for exclusively manual video annotation. The annotation scheme is generic and customizable. Customization can be done by specifying a set of attribute-value pairs which are used to attach the metadata. The structuring possibilities are simple definitions of annotat-

19 N. Dimitrova, Multimedia Content Analysis: The Next Wave, *Proc. of International Conference on Image and Video Retrieval*, Urbana-Champaign, July 2003, 9–18. C. G. M. Snoek and M. Worring, Multimodal Video Indexing: A Review of the State-of-the-art, *Multimedia Tools and Applications*, 25(1), Jan. 2005, 5–35.

20 C. G. M. Snoek and M. Worring, Multimodal Video Indexing.

21 The Deconstructor: An Online Film Analysis System, http://ccnmtl.columbia.edu/projects/feature_pages/133-deconstructor.pdf.

22 http://www.dfki.de/~kipp/anvil.

able frame sequences. Its original purpose was to annotate gesture and speech semantics in videos. To give an overview of the annotations in a specific time interval, *Anvil* has a so-called annotation board. This sub-window shows tracks of different data types like sound, spoken words, gesture descriptions, etc. The *Anvil annotation tool* aims at annotating parts of a video.[23] The weaknesses of *Anvil*, however, are its limited structuring possibilities and the need for explicit types for the annotations.

The *M-OntoMat-Annotizer*[24] is a public semantic annotation tool that was developed in the context of the *aceMedia* project.[25] Basically, it enables the user to attach metadata to videos or images. The basic idea of the tool is to extract low-level MPEG-7 descriptors and link them automatically to ontologies and semantic annotations in order to annotate high level semantics. Descriptors in the MPEG-7 standard are evaluations of video frames or individual images.[26] An example would be the dominant color in a region of interest in a frame or an image. At the time we evaluated the *M-OntoMat-Annotizer*, it did not support the manual structuring of videos. Instead, it concentrated on generating descriptors of images or frames. These descriptors are analyzed and associated with ontologies by an internal knowledge base which is an automated process. The *M-OntoMat-Annotizer* has a very different approach to video annotation. Structuring features in this tool are barely useful because it generates the structure by linking the descriptors with matching ontologies. The main feature is the automatic linkage between technical descriptions of the video and ontologies.

The *VideoAnnEx annotation tool*[27] allows the user to annotate shots in a video. The annotation data is stored in an MPEG-7 file. Each shot in the video can be annotated with static scene descriptions, key object descriptions, event descriptions, or other lexicon sets of descriptions. This restricts the annotation possibilities to the content of the lexica but keeps the annotations simple and consistent.[28] The *VideoAnnEx* tool takes an MPEG video and an optional MPEG-7 annotation file as its input. In case the annotation file is not found,

23 M. Kipp, ANVIL – A Generic Annotation Tool for Multimodal Dialogue, *Proceedings Eurospeech*, 2001. M. Kipp, *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*, Boca Raton, Dissertation.com, 2004.

24 http://acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html.

25 http://acemedia.org/aceMedia.

26 K. Petridis, S. Bloehdorn, C. Saathoff, N. Simou, S. Dasiopoulou, V. Tzouvaras, S. Handschuh, Y. Avrithis, Y. Kompatsiaris, and S. Staab, Knowledge Representation and Semantic Annotation of Multimedia Content, *IEE Proceedings – Vision, Image, and Signal Processing*, June 2006, Volume 153, Issue 3, 255–262.

27 http://www.research.ibm.com/VideoAnnEx/usermanual.html.

28 C.-Y. Lin, B. Tseng and J. Smith, VideoAnnEx: IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning, *Proc. of ICME*, Baltimore, July 2003.

VideoAnnEx creates the file by performing shot boundary detection. The shots can also be detected by the *IBM CueVideo Shot Detection Toolkit.*

From this review of the state-of-the-art, the necessity for a specific tool tailored to the needs of film analysts in terms of locating/analyzing religious references can be directly derived. According to our knowledge, none of the existing tools supports a flexible spatio-temporal annotation feature in combination with an effective navigation interface and sufficient automatic analysis features to support the analyst in his or her work.

3. Semantic Video Annotation Tool (SVAT)

In this chapter, we introduce our Semantic Video Annotation Tool (SVAT) application which supports spatio-temporal annotation. The sections below describe a typical annotation workflow from the user's perspective. This workflow typically starts with preparatory steps (i.e. preprocessing) where several features needed in subsequent steps are fully automatically extracted. This is described in section 3.1. In the second part of this chapter (section 3.2), the manual annotation process is described which utilizes the automatically extracted features.

3.1 Preparatory Steps and Automatic Annotation Assistance

In a first preparatory step of the annotation workflow, the global description of the source video data has to be specified and an automatic analysis process has to be started. A separate tool, the Media Analyze Tool (see Figure 1), has been developed for this. During the automatic content analysis, the characteristic camera motion, shot boundaries including dissolves, relevant key frames, several image similarity features, and the features which are required for locating similar image regions in the video are extracted. The automatically extracted metadata is described in detail in the following sections. The result is a first metadata description of the video which is saved in the XML metadata standard MPEG-7. The content analysis is a very time-consuming process. Therefore, the automatic analysis can be started for several videos and then the process can run over night.

3.1.1 Shot Boundary Detection

Shots are the basic building blocks of the visual modality of a video. Shot boundary detection algorithms identify both abrupt (hard cuts) and gradual transitions (such as dissolves, fades, wipes, etc.) and produce a decomposition of the timeline into shots (shot list). As shot boundaries are the natural limits of most visual features (such as camera movements, objects), their detection is

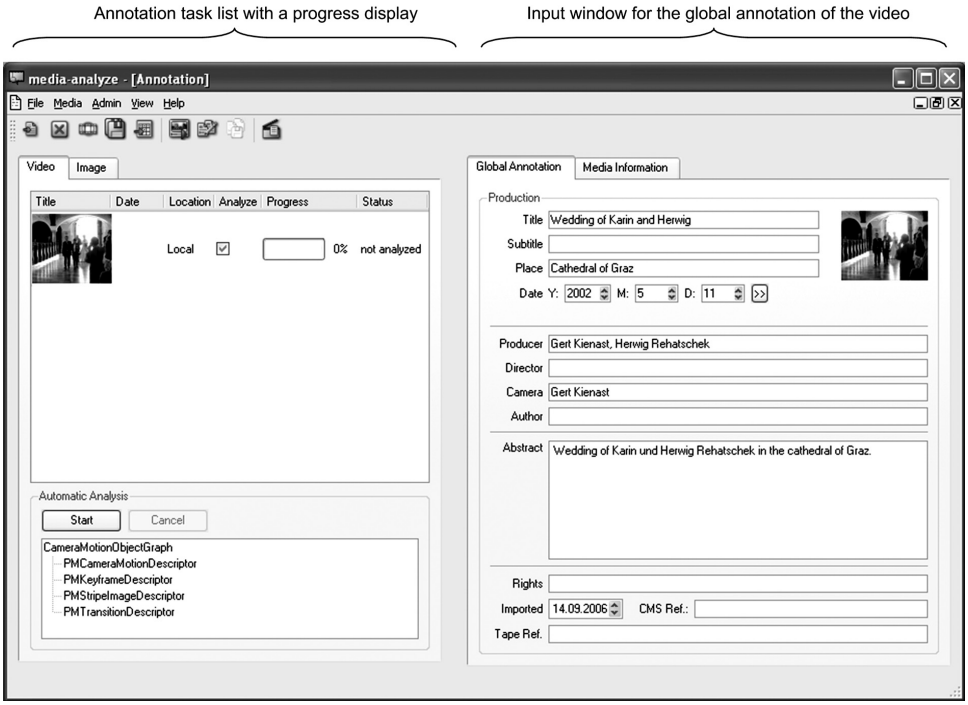

Figure 1: Media-Analyze Tool for the automatic extraction of metadata

an important prerequisite for further automatic content analysis. The visualization of the shot supports the user in efficiently navigating through the video and accessing certain time segments within the video. Shots are also an important basis for manual annotation, as many annotations are shot-based and knowledge about the shot structure facilitates building higher-level content structures such as scenes, sequences or chapters. In our semantic video annotator, an improved version of the algorithm described by Bailer et. al. is used.[31]

3.1.2 Key Frame Extraction

Key frame extraction tools identify a number of frames in each shot, which are used as representative proxies of the shot's content and support effective navigation within the video. The selection of key frame positions is based on the visual activity in the visual essence, i.e. the more object and/or camera motion, the shorter the time interval between two key frame positions. At least one key frame is extracted at the beginning of a shot. Key frames are an important means for efficient representation of the visual content of a shot and support navigation

31 W. Bailer, H. Mayer, H. Neuschmied, W. Haas, M. Lux, W. Klieber, Content-Based Video Retrieval and Summarization Using MPEG-7, *Proc. of Internet Imaging V*, San Jose, Jan. 2004, 1–12.

in the content. The extracted key frames are visualized in a special view of the SVAT which is located near the bottom as shown in Figure 4.

3.1.3 Stripe Image

Stripe images are spatio-temporal representations of the visual content and hence an effective way to navigate within video footage. A stripe image is created by extracting one column from each of the frames of the video and putting them together in temporal order. This visualization clearly shows shot boundaries as discontinuities, and allows a user to easily track camera motion and movements of larger objects. The extracted stripe images are visualized in the stripe image view of the SVAT tool (as can be see at the bottom of Figure 4).

3.1.5 Object Recognition and Image Retrieval in Video

Automatic object and image retrieval in video material is an essential and necessary part of any annotation tool because manually searching for objects in a video is a time-consuming and annoying task. Supporting semi-automatic annotation, this computer-aided object retrieval should be easy to use, very fast and the results should be as accurate as possible. The retrieved images should be presented in a ranked list of key frames just as web search engines list their results in terms of relevance. After verification of the result whereby very little user interaction is necessary, the correctly found key frames can be annotated automatically. The resulting object recognition and video retrieval module was implemented as a plug-in for the SVAT tool.

The retrieval of an object in a video is a challenging task because an object's visual appearance may vary within the video due to changes in perspective or

32 W. Bailer, P. Schallauer, G. Thallinger, Joanneum Research at TRECVID 2005 – Camera Motion Detection, *Proc. of TRECVID Workshop*, Gaithersburg, Nov. 2005.

lighting, or the object may be partially occluded. To overcome these problems, several approaches have been developed based on a weak segmentation of the image instead of segmenting the image semantically in object foreground and background. Currently this 'local approach' to object recognition is very popular. The basic idea is to 1. detect a set of discriminative interest points (key points); 2. describe the local image region around the key point in a (possibly) invariant manner using a so-called descriptor; 3. use the set of descriptors for matching and recognition (basically a voting algorithm that counts the number of matches).

The first step commonly taken for video processing is to split a given video into shots. The advantage of shot detection is to guarantee that at least one frame per shot is indexed and therefore objects can also be located in very short shots. We suggest defining those frames with the highest number of detected low-level features within every shot as key frames and to index them for later

33 D. G. Lowe, *Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision,* 60, 2 (2004), 91–110.

34 J. Matas, O. Chum, M. Urban, T. Pajdla, Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In: *Proceedings of the British Machine Vision Conference*, 2002, 384–393.

35 R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press, 1999.

36 D. Nistér and H. Stewénius, Scalable Recognition with a Vocabulary Tree. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2, 2006, 2161–2168.

retrieval instead of constantly using one frame per second to force better query results (these key frames are solely used for object detection purposes and not for the navigation purposes described in section 3.1.2). When a new key frame of the movie is observed, each stable descriptor of the frame is assigned to the nearest visual word. The benefit of this casting is that matches are effectively pre-computed so that at run-time, frames can be retrieved immediately by matching the features with the visual words in the vocabulary. Because very few visual words are common with many entries and many visual words only have a single entry, the application of a so-called 'stop list' is highly recommended, where those visual words in the vocabulary are deleted. This implies a reduction of the vocabulary size without any loss of performance in terms of accuracy.

In the retrieval stage, the defined region of interest contains low-level features which are used to match the visual words in the vocabulary. Every key frame which is retrieved by a successful match gets a vote and is inserted into the hit list. This hit list of key frames is finally ranked by the number of votes and presented to the user.

Figure 2 illustrates the implemented application. The image on the left shows the video player with which the user can navigate through the video and mark interesting objects with a rectangle, the so-called 'region of interest'. After the selection, the search process can be started and within a few seconds all key frames found are listed in a ranked hit list as shown in the right image. If an image that was found does not match the requirements of the user, its checkbox can be unmarked. After verification, all key frames marked as correct results can be annotated in the video by the application automatically. It is also possible

Figure 2: Left Image: Video player with object to search for in the region of interest. Right Image: Hit list with retrieved key frames. The correct key frames are marked for later annotation.

to jump directly from a key frame in the hit list to the position in the video player to perform another search and to come across more similar occurrences of this object.

3.2 Manual Video Annotation

As soon as the automatic analysis of a video is completed, the produced metadata description can be displayed, edited and extended by the Media-Annotation tool (see Figure 4). This tool has a number of views which enables quick and easy navigation in the video. Through the key frames and the stripe image, the user is provided with a quick overview of the video content. There are two time lines, one for the whole video time and one which shows only a selected time period (time zoom). In the time lines, the shot boundaries and the dissolves are displayed and they can also be edited. There is also the option of structuring the video depending on the video content. For example, shots can be grouped into scenes; scenes can be combined to build chapters and so on. This structure yields a table of contents and is displayed by a separate view (see Figure 4). Depending on the selected structural element, different textual annotations are possible. These are, for example, the title of the structural element, content description, remarks, and specifications about time, location, and persons. At the shot level, shooting settings like camera motion, camera angle, or view size can be documented (see Figure 3).

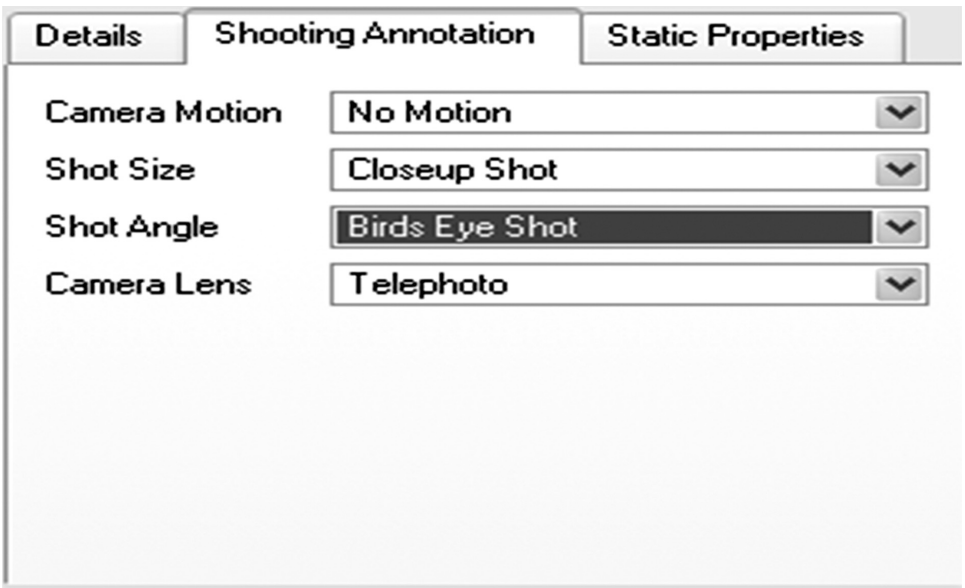

Figure 3: Possible annotation of the shooting settings

The integrated video player has drawing functionalities for the annotation of regions or objects. Once an image region is specified (see white rectangle at the left side in Figure 4), it is possible to start an automatic search for a similar

Figure 4: Semantic Video Annotation Tool (SVAT)

region in other frames of the video. For this, the pre-calculated SIFT feature values are used (see chapter 3.1.5). Objects which do not change their visual appearance and shot locations (background regions) can be redetected. The search result is a list of key frames which include the detected regions. This list can be manually edited and is used to assign textual annotations of objects, persons, locations, etc. (which have to be specified only once) to several shots, scenes or any other structural unit.

4. Conclusions and Outlook

We have presented the video and film annotation tool SVAT, which enables film analysts to efficiently annotate video content. In comparison to existing annotation tools, which in most cases only allow global annotations, SVAT allows for temporal and spatial annotations. Hence, objects moving in time (e.g. a person moving from one point to another) can be annotated as well as single objects within one frame. SVAT also supports the transformation of spatial into temporal annotations by means of automatic tracking. The user marks a certain object and SVAT automatically tracks the object forwards and backward, thereby generating an automatic annotation of it. Moreover, the tool offers innovative mechanisms for navigation by several configurable views such as a shot, key, and stripe image view. With regard to structuring, the user can individually and hierarchically group parts of the film (e.g. shots, scenes) together, starting from a frame level.

With respect to object recognition, we presented an approach based on Difference-of-Gaussian key points and computed SIFT descriptors as local low-level features for every frame in the video. These key point locations are very precise, repeatable, and also highly scalable. Therefore, it is possible to add more local feature detectors and descriptors. The big advantages of our specific approach are that the user does not waste time with interactive training and it is possible to preprocess every video without any technical knowledge. After the offline preprocessing task, which may take a few hours, the video is immediately ready for use. There are no limitations on the underlying video format and it is possible to use black-and-white videos should any color features be required.

The presented tools enable and facilitate the description, structuring, and analysis of a video. Of course, the result should be reusable for further analysis steps with respect to the video itself but also in relation to other annotated videos. As for the annotation tool, additional functionalities are required for generating statistics and reports. A database for managing the video metadata as well as a database search tool have already been implemented. The extension of these tools is planned, with specific functionalities which should facilitate the analysis and comparison of different videos.

In order to facilitate the interpretation of symbols in media products, a database is developed which should link traditional symbols of Christian iconography with those used in media. Keeping in mind the technical advancements in object recognition, as well as its current technical limitations, we will consider it as important to develop an algorithm that is able to link objects in videos to those database entries.

Acknowledgments

Parts of this work were funded by the Austrian Science Fund (FWF) and the European Union as part of the research projects PrestoSpace (IST-FP6-507336) and GMF4iTV (IST-2001-34861).