Efficient Semantic Video Annotation by Object and Shot Re-Detection

Peter Schallauer
JOANNEUM RESEARCH
Steyrergasse 17
8010 Graz
peter.schallauer@joanneum.at

Sandra Ober
TU Graz
Inffeldgasse 16/2
8010 Graz
ober@icg.tu-graz.ac.at

Helmut Neuschmied
JOANNEUM RESEARCH
Steyrergasse 17
8010 Graz
helmut.neuschmied@joanneum.at

ABSTRACT

Manual video annotation on shot and on object level is a very time consuming and therefore cost intensive task. Automatic object and shot re-detection is one step forward in order to provide a cost efficient solution for temporally detailed video annotation. In this demonstration a tool will be shown which integrates novel video visualisation, navigation and interactive object re-detection techniques. Automatically re-detected objects and shots can be annotated by reusing operator annotations in a semi-automatic way. Object and shot re-detection capabilities are demonstrated on a variety of content, e.g. anchor person re-detection in news content and setting and re-take detection in rushes content.

Keywords

Video, semantic, interactive, annotation, object re-detection.

Categories and Subject Descriptors

H.5.1 Multimedia Information Systems.

1. INTRODUCTION

Many techniques exist to analyse, annotate and further interpret multimedia content. However, nowadays these techniques are still limited with respect to the information they can automatically generate, and therefore it is important to provide tools that assist users in correcting, refining, and creating annotations in a semi-automatic way. Manual video annotation on shot and on object level is a very time consuming and therefore cost intensive task. Section 2 presents a tool which aims at efficient manual annotation by automatic video pre-processing. Section 3 describes a SIFT based object and shot re-detection technique, which is utilised for semantic video annotation presented in section 4.

2. SEMANTIC VIDEO ANNOTATION SUITE

The Semantic Video Annotation Suite (SVAS) consists of two tools. A pre-processing tool, called *Media Analyze* Tool, performs fully automatically most of the computational work of video analysis, and stores the results in a MPEG-7 data base. With the *Annotation Tool*, the user selects video objects interactively and relates them to semantic descriptions. In order to reduce the annotation effort, the *Annotation Tool* contains automatic object and shot re-detection functionalities, which uses the preprocessed data to reduce the computation time. All produced metadata is stored in the ISO standard MPEG-7 compliant to the DAVP [1] profile.

The *Media Analyze* Tool allows for fully automatic content analysis, metadata for video navigation and structuring is generated. Shots, key-frames, stripe images and image features required for object and shot re-detection are extracted. The analysis process can be started for several videos and then the process can run over night.

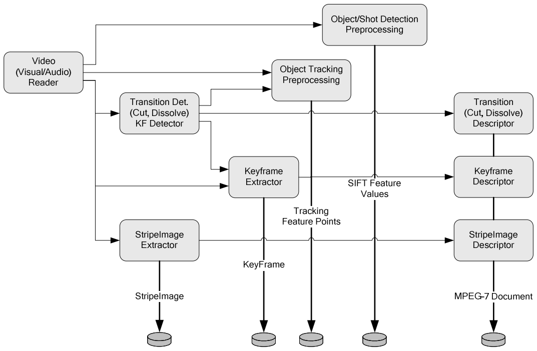
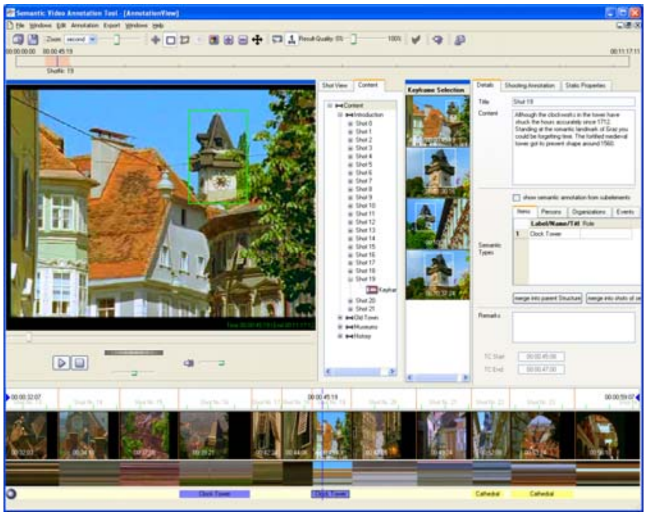

Object/Shot Detection Preprocessing
Video (Visual/Audio) Reader
Object Tracking Preprocessing
Transition Det. (Cut, Dissolve) KF Detector
Transition (Cut, Dissolve) Descriptor
Keyframe Extractor
SIFT Feature Values
Keyframe Descriptor
Tracking Feature Points
StripeImage Extractor
StripeImage Descriptor
KeyFrame
StripeImage
MPEG-7 Document

Figure 1 Analyses module graph of the *Media Analyze* Tool

The core part of this program is a module framework. The framework serves as an execution environment for analysis modules. The modules are interconnected by the framework to constitute a so-called module graph (see Figure 1). The module graph is defined by a XML file.

As soon as the automatic analysis of a video is completed, the produced metadata description can be displayed, edited and extended by the *Annotation Tool* (see Figure 2).

This tool has a number of views which enable fast and easy navigation in the video. Through the key frames and the stripe image, one gets 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.



Title: Shot 19

Content: Although the clockworks in the tower have struck the hours accurately since 1712. Standing at the romantic landmark of Graz you could be forgetting time. The fortified medieval tower got its present shape around 1560.

☐ show semantic annotation from subelements

Semantic Types:

	Label/Name/Titl	Role
1	Clock Tower	

Remarks:

TC Start: 00:00:45:00

TC End: 00:00:47:00

Figure 2 User interface of the video *Annotation Tool*

The Annotation Tool also allows the annotation of the temporal video structure in a flexible way. Shots, scenes and chapters can be arranged hierarchically. The user can decide on the needed level of temporal abstraction. On each level (shot, scene, chapter and entire video) the occurrence of semantic concepts (who, what, where, when, why,...) can be annotated as free text or by usage of controlled vocabulary originated from classification schemes and semantic knowledge bases.

3. OBJECT RE-DETECTION

Our approach for object and shot re-detection is inspired by *Video Google* [2] and consists mainly of these three components:

- The representation of every frame by a set of viewpoint invariant key point descriptors
- The use of consecutive frames to improve descriptors quality
- Vector quantization of the obtained descriptors

In our application we detect Difference-of-Gaussian key points and compute SIFT (Scale Invariant Feature Transform) descriptors introduced by Lowe [3] as local low-level features for every frame in the video. Every descriptor is aggregated from more contiguous frames to reduce noise. Any key point that does not survive for more than five frames is refused to reject unstable descriptors. To each descriptor a scale and orientation value is assigned. The scale is obtained by the size and the orientation is calculated from the local gradient directions of the key point.

The next step is to setup a visual vocabulary. The objective is to vector quantize the descriptors into clusters. Vector quantizing brings a huge computational advantage because descriptors in the same clusters are considered as matched and no further matching on individual descriptors is required. Instead of clustering all descriptors simultaneously, which is impossible only a subset is selected. Once the visual words are defined, all stable descriptors of a key frame are assigned to a visual word according to the nearest cluster center. The frequency of occurrences of every visual word across the whole video (or database) is measured and the top and bottom 5% visual words are stopped. This step is inspired by the stop list used in text retrieval applications, where very common and uncommon words are discarded.

Every video is represented as a set of key frames and each is represented by the visual words it contains. The corresponding positions of visual words within a frame are stored in an Inverted File Structure (IFS) and is organized comparable to an ideal book index. In our case the IFS has an entry for each visual word, which stores all the matches.

In a text search process the ranking is increased for documents where the searched words appear close together in the retrieval texts. In our case the matched descriptors in the retrieved frames should have a similar spatial arrangement to those of the outlined region in the query image. For this reason a hypothesis of the object pose is calculated from the scale and orientation value of matched descriptors. The matched descriptors, which have a correlation in their object pose hypothesis, are grouped. Each of these groups represents an object re-detection result. The ranking of the result is determined by the number of matched descriptors.

The object detection rates strongly depend on the textural information of the image regions and on the size of object appearances. Objects that contain much textural information can be recognized with a higher reliability because they are represented by a large number of stable descriptors. Best results can be achieved with non-moving, rigid and planar objects of adequate appearance size and quality.

4. OBJECT AND SHOT ANNOTATION

For object annotation the video player integrated in the *Annotation Tool* has drawing functionalities. An image region can be specified for a certain object by drawing a rectangle or a polygon. For shot re-detection the entire image region is specified. Once the object location is defined (see green rectangle in Figure 2) it is possible to start an automatic matching for similar objects respectively shots in the entire video. The result of the object re-detection is displayed in a separate key-frame view. In each of the displayed key-frames the found object location is indicated. The manual annotation of a certain object can be copied by one mouse click to all matching objects within the video, thus reducing massively the manual annotation time required.

5. CONCLUSIONS

The work intensive task of manual video annotation on object level can be heavily decreased by automatic object and shot re-detection techniques. The SVAS tools, available as free use demonstrator software at [4], are implementing this approach.

6. ACKNOWLEDGMENTS

This research was supported by the European Commission under contract FP6-027026, K-Space.

7. REFERENCES

[1] Bailer, W. and Schallauer, P. 2006. The Detailed Audio-visual Profile: Enabling Interoperability between MPEG-7 Based Systems. In Proceedings of IEEE MMM ‘06.

[2] Sivic, J. and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In Proceedings of ICCV ’03, pp. 1470-1477.

[3] Lowe, D. G. 2004, Distinctive Image Features from Scale-Invariant Keypoints. In IJCV ’04, 60(2), pp. 91-110.

[4] Semantic Video Annotation Suite Demonstrator Software: http://www.joanneum.at/en/fb2/iis/products-solutions-services/semantic-video-annotation.html