Ralph Ewerth, Markus Mühling, Thilo Stadelmann,
Julinda Gllavata, Manfred Grauer, Bernd Freisleben

Videana: A Software Toolkit for Scientific Film Studies

Abstract

Within the research project "Methods and Tools for Computer-Assisted Media Analysis" funded by Deutsche Forschungsgemeinschaft, we have developed the software toolkit *Videana* to relieve media scholars from the time-consuming task of annotating videos and films manually. In this paper, we present the automatic analysis tools and the graphical user interface (GUI) of *Videana*. The following automatic video content analysis approaches are part of *Videana*: shot boundary detection, camera motion estimation, detection and recognition of superimposed text, detection and recognition of faces in a video, and audio segmentation. The GUI of *Videana* allows the user to subsequently correct erroneous detection results and to insert user-defined comments or keywords at the shot level. Furthermore, several research applications of *Videana* are discussed. Finally, experimental results are presented for the content analysis approaches and compared to the quality of human annotations.

Introduction

The research project "Methods and Tools for Computer-Assisted Media Analysis" (MT) of the collaborative research center "Media Upheavals" develops (a) the database system *Mediana* to allow media scholars to manage arbitrary textual and audio-visual data objects and supports related research work flows, and (b) the software toolkit *Videana* as part of *Mediana* which includes computer-assisted methods to support the scholarly analysis of audio-visual material, in particular images and videos.

In this paper, we will discuss how far academic film studies can be supported by *Videana* in a quantitative manner. Clearly, the interpretation of audio-visual scenes will be reserved exclusively to humans for a conceivable time. Nonetheless, computers can disburden media scholars from typically very time-consuming manual annotation tasks. In particular, the quantitative analysis of the following elements of film and video composition can be supported: montage of shots (e.g., cut frequency), camera motion, a description of dis-

played scene content, mainly with respect to faces and superimposed text, and audio information.

Korte (2001) describes several elements of cinematic composition, some basic elements of shot and sequence protocols as well as several types of visualization, such as graphics displaying shot composition, shot and sequence protocols and cut frequency diagrams. Some years ago, researchers have developed computer systems which support the task of creating shot protocols and simplify the generation of visualizations (e.g. scene and sequence diagrams, cut frequency diagrams). For example, the system *filmprot* was developed at the University of Marburg (Institute for Media, Giesenfeld 1991), while Korte developed *CNfA* ("Computergestützte Notation filmischer Abläufe", i.e. "Computer-based notation of cinematic episodes"; Korte 1992, 1994) at the University for Visual Arts of Braunschweig. However, these systems worked only in conjunction with particular analog video recorders which are not available anymore. The software *Akira (University of Mannheim,* Kloepfer*),* the software *VideoAS* (University of Jena, Olbrecht/Woelke), and the website *CineMetrics.lv* belong to the more recent developments for the purpose of annotating (digital) videos but they do not include tools for automatic video content analysis.[1]

Finally, it should be mentioned that media content analysis is not only of interest for purposes of media studies. The proliferation of media data is rapidly increasing, e.g. if one considers the popularity of MP3 music files, digital photo collections of home users, digital videos, web-based video databases (e.g. youTube.com) and IPTV (Internet Protocol Television, e.g. *Joost:* www.joost.com). Hence, it is obvious that the need for efficient search operations in large media databases is growing accordingly. Anticipating these recent developments, efficient content-based search in media databases has been a field of extensive research since the middle of the 1990s.

Videana: A Software Toolkit for Video Analysis

As mentioned above, the big advantage of computer assistance is the automation of formal and compute-intensive analysis tasks. For example, these are tasks such as the temporal segmentation of a video into shots, identification of the kind of montage of subsequent shots (cut, dissolve, fade, etc.), finding and recognizing superimposed text, recognition of camera and object motion, recognition of camera distance, information about the presence of actors, recognition of audio events etc.

1 See the papers by Tsivian and Kloepfer in this volume for a description of their applications.

Figure 1: The main window of *Videana*. On the left side, there is a window for playing a video. There are two timelines at the bottom which visualize the analysis results for the temporal segmentation of the video into shots as well as for face detections. The vertical lines in the *Cuts* timeline represent cuts (abrupt shot changes), the colored (here: grayish) areas in the timeline *Faces* mark the sequences where a frontal face appears. Two timelines are presented for each kind of event: the upper one represents the total duration of a video, whereas the lower one zooms into a certain time period which can be selected in the upper timeline and is surrounded by a rectangle. Further timelines are displayed in case that the corresponding analysis results or user annotations are available for the related events of camera motion, superimposed text or audio events. On the right side, the temporal segmentation is presented in another way. Single shots are represented by three frames (beginning, middle, and end frame of a shot). By a mouse click on an icon, the related video frame is accessible directly, while a double click starts playing the video from this position.

Up to now, the following automatic video content analysis algorithms are integrated in *Videana*: Shot boundary detection, text detection and recognition (video OCR: optical character recognition), estimation of camera motion, face detection, and audio segmentation which segments the video into sequences of silence, speech, music and background noise. Based on a plug-in approach, any type of analysis algorithm can be updated, exchanged or removed easily. The graphical user interface (GUI) of *Videana* allows users to play videos and to access particular video frames. Furthermore, the GUI allows users to manually correct erroneous analysis results.

As soon as a temporal segmentation of a video has been obtained, an icon is created for the first, the middle and the last frame of each shot. These icons are displayed to the user in the shot list view (see Figure 1). Such a view is also possible for scene segmentation but currently this segmentation has to be provided by the user manually. *Videana* offers functions to automatically generate diagrams with respect to brightness changes and cut frequencies for a video. Figure 2 shows a cut frequency diagram for a 30-minute movie sequence. The results of the different analysis algorithms are visualized in separate timelines: *cuts*, *text*, *camera*, *face*, and *audio*. The user can insert arbitrary comments and keywords for particular events and shots. All these metadata, generated either automatically or manually, can be saved in an MPEG-7 ("Multimedia Data Description Interface", Martinez 2002) XML file. The MPEG-7 standard formalizes the representation of metadata for audio-visual objects and establishes a basis for data exchange between different multimedia applications.

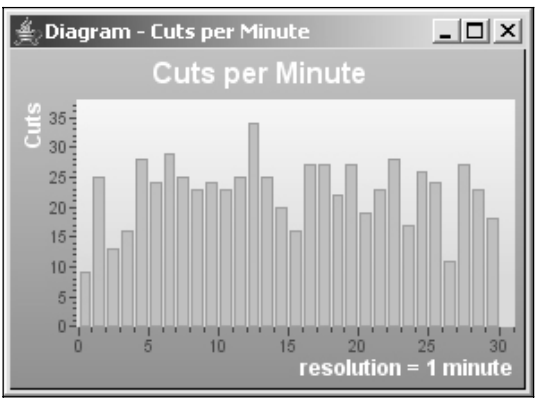

Figure 2. A cut frequency diagram generated by *Videana* for a 30-minute film sequence.

Shot Boundary Detection

One of the most important tasks in digital video analysis is the segmentation of a video sequence into its fundamental units, the shots. A shot is generally understood as an audio-visual sequence recorded continuously without any interruption. The transitions between shots can be abrupt or gradual; abrupt transitions are called cuts, gradual transitions are the results of chromatic or spatial editing effects, such as fade in/out, dissolves or wipes.

Since the beginning of the 1990s, a huge number of segmentation approaches have been suggested. Widely known approaches were suggested by Yeo/Liu (1995), Hanjalic (2002), or Bescos (2004), particularly for cut detec-

tion. To detect gradual transitions, general approaches as well as specialized detectors for certain effects have been developed (e.g. Hanjalic 2002 for dissolves, Truong et al 2000 for fade in/out). Many approaches for cut detection are based on the comparison of two consecutive frames. More recent approaches (Tahaghoghi et al 2005, Yuan et al 2005) compare all images within a short time window with each other to get more robust results. In the TRECVID evaluation series[2] such approaches could achieve the best recognition rates in 2005: 95% of the cuts were found (recognition rate, "recall"), and 95% of all positions reported by these detectors were indeed cuts ("precision" of the result). The approach developed by the authors (Ewerth/Freisleben 2004) belonged to the top five approaches, which achieved a recognition rate as well as a precision of at least 90%. Twenty-one institutes from all over the world participated in this study. The detection of gradual transitions has not yet reached this quality. Here, the recognition rate and the precision of the best approaches (Amir et al 2005; Yuan et al 2005) achieve approximately 80%.

Camera Motion Recognition

From an aesthetical point of view, camera motion is often used as an expressive element in film production. Video compression formats like MPEG-1 or MPEG-2 exploit the large temporal redundancy in videos for data compression and thus support motion estimation based on pixel blocks for consecutive video frames. The runtime for the extraction of such motion vectors is very low compared to the decoding of a whole image and the calculation of the optical flow field (calculation of motion for each pixel). Although the use of MPEG motion vectors improves runtime performance, a big part of these vectors is often "noisy" and thus not optimal in the sense of a motion description. Based on these observations, we have developed an approach (Ewerth/Schwalb/Tessmann/Freisleben 2004) which uses MPEG motion vectors for calculating the camera parameters. The "unreliable" motion vectors of a vector field are removed by an effective method in a preprocessing step, called "outlier removal". The parameters of a 3D camera model are estimated by means of these remaining motion vectors using the Nelder-Mead algorithm for solving the minimization problem. The used model has the advantage that it basically allows the distinction between camera translation and camera rotation in the corresponding direction. Experiments with synthetic video sequences

2 TREC, the Text Retrieval Conference, conducted a video retrieval evaluation for the first time in 2001. Since 2003, there is a separate annual Video Retrieval Evaluation Workshop called TRECVID; see also: www-nlpir.nist.gov/projects/ t01v.

showed that outlier removal leads to clearly better results. For zoom-in and zoom-out, a recognition rate and a precision of 99% (98% and 94% without outlier removal) could be achieved and the results for rotation around the z-axis could be improved from 86% to 95% (recognition rate) and from 75% to 89% (precision). We participated in the TRECVID evaluation 2005 using the described system. Overall, twelve institutes participated in this "low-level-feature detection task" concerning camera motion. The evaluation required the analysis of 140 news videos in total with a respective duration of 30 to 60 minutes. The submitted results should include all camera shots which contain horizontal, vertical camera movement or zoom (in/out). For the purpose of evaluation, the organizers finally selected approximately 2000 shots from the 140 videos obviously containing (or obviously not containing, respectively) camera motion or zoom. Besides achieving good results for the recognition of horizontal camera movement ("pan": 76% recognition rate, 92% precision), our system reached the second-best result regarding vertical movement ("tilt": 72% recognition rate, 96% precision) and the best result with respect to zoom recognition (89% recognition rate, 93% precision).

Detection and Recognition of Superimposed Text

Superimposed text often hints at the content of an image. In news broadcasts, for example, the text is closely related to the current report, and in silent movies it is used to complement the screen action with intertitles. Involved algorithms can be distinguished by their objective, whether it is text detection, localization or tracking (in videos), text segmentation (also called text extraction) or text recognition (Jung et al 2004). A text detector answers the question whether there is any text in an image or shot, and where it is. Then, text segmentation crops localized text out of the image to yield black letters on a white background. This step is necessary to feed the result into an OCR program, which transforms the image into machine-readable text. A non-uniform background would impair this process. Exemplary results of these three stages are depicted in Figure 3.

Automatic optical character recognition (OCR, normally on scanned text pages) has been a research topic since decades, and text detection, segmentation and finally recognition in images and videos has been investigated for more than 10 years now. This led to a plethora of methods that are surveyed by Jung et al (2004). The work conducted in the authors' workgroup includes proposals for text detection (Gllavata/Ewerth/Freisleben 2004a) and text segmentation (Gllavata/Ewerth/Stefi/Freisleben 2004; Gllavata/Freisleben 2005) as well as a method for tracking moving text across several video frames

(Gllavata/Ewerth/Freisleben 2004b). The proposed text segmentation method was able to boost the word (character) recognition rate from 62% to 79% (76% to 91%) on a set of test images (Gllavata/Freisleben 2005). Recently, the Tesseract OCR engine (Vincent 2006) has been integrated into *Videana*, such that the software is now able to annotate shots with localized and recognized words automatically.

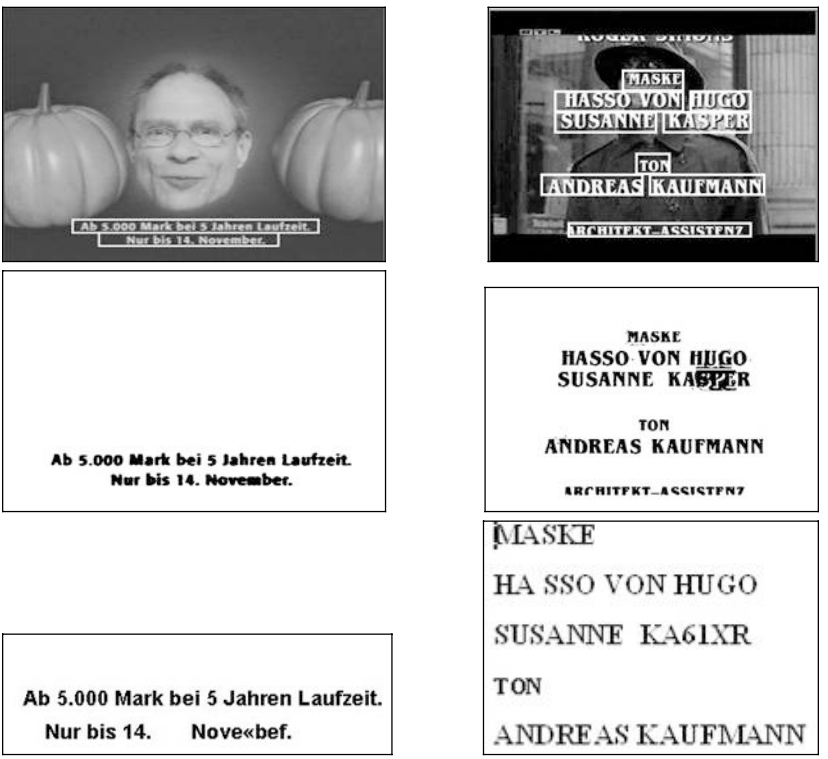

Figure 3: The images on the top row show the result of text localization. The middle row depicts the results of the text segmentation process, in which the background has been removed and the text has been marked black. In the last row, the result of the recognition software (OCR) is shown.

Detection and Recognition of Faces

Face processing in images and videos is composed of face detection and face recognition. Yang et al (2002) give a comprehensive survey on the detection problem, while face recognition is surveyed by Zhao et al (2003). Face detection can be viewed similar to text detection: The detector decides whether there are faces in an image or shot, and typically solves the localization prob-

lem, too. *Videana* applies the method of Viola and Jones (2004) from the Intel Open Computer Vision Library (OpenCV) to detect frontal faces. It yields a detection rate of 92.1% on relevant standard test sets (containing 130 images with 507 faces overall) with 50 false positives (Viola and Jones 2004). In face recognition, two different scenarios can be distinguished:

- Identification: the system has to recognize the identity of a given face image by comparing it to known faces in its database, or to reject it as unknown.
- Verification: the system judges whether a given face image fits a given identity claim based on its database.

The 2002 Face Recognition Vendor Test (FRVT) has shown (Phillips et al 2002) that current face recognition technologies are able to achieve recognition rates higher than 90% under certain conditions. The identification (or recognition) rate is the percentage of faces that could be matched correctly with a known face in the database. If a face could not be recognized, two kinds of errors are possible: False alarms (or false accept), meaning that a face was falsely recognized as a wrong (known) face; and false rejects, meaning that a known face was not recognized as known at all (Phillips et al 2007). Summarizing, the study of Phillips et al (2002) drew the following conclusions:

- Best systems reached identification rates of 90% (false alarm rate 1%) for indoor images. For only 0.1% false alarms, the identification rate was 80%.
- The better systems were not sensitive to illumination for normal indoor images.
- 3D models improved the results, by morphing the head's pose to a frontal view.
- The identification of faces in outdoor images did not work satisfactorily: 50% identification rate at a false alarm rate of 1%.
- There was no difference in recognition rate whether single still images or video sequences were used as source material.
- It was more difficult to recognize younger faces than older faces.
- The recognition rate for male faces was higher than that for female faces.
- The recognition rate dropped linearly with the logarithm of the database size (number of persons).

The 2006 FRVT study (Phillips et al 2007) assessed the development of industry-strength face recognition technology since 2002. It also covered iris recog-

nition (Iris Challenge Evaluation 2006) and investigated 3D face data and high resolution images in addition to the previous test. Several improvements over the older results are reported:

- The identification rate was improved considerably, lowering the false rejection rate (having 0.1% false alarms) by a factor of 4 to 6, depending on the actual algorithm.
- Under improved illumination conditions and using very high resolution images, the identification rate reached 99% (at 0.1% false alarms), corresponding to a reduction of false rejections by a factor of 20.
- Handling of uncontrolled conditions was improved, reaching the controlled identification rates of 80% (at 0.1% false alarms) of the 2002 candidates.
- Interestingly, the top algorithms were able to match or even do better than human face recognition performance on unfamiliar faces under illumination changes.

These results show the state of the art in face recognition technology when one can control the circumstances under which face images are taken. Taking into account the lower resolution of videos and that the recording environment might be arbitrary in videos, these results also suggest that it is difficult to index a video by appearing faces. This and the better exploitation of the large amount of single still images in a video sequence is an area for future research.

Although a specific face identification system may be useful for media research purposes (i.e. to answer questions like "in which shots did a given person appear?"), *Videana* currently contains a *general* person recognition system yielding an index of appearing persons over time for any video (Ewerth/Mühling/Freisleben 2006). Its single prerequisite is just a given segmentation of the video into shots or, optionally, scenes. The result is a set of appearing persons, and for each a list of shots in which he or she appeared. In principle, this system is able to detect and recognize both frontal and profile faces, but the currently employed OpenCV detector is not technically mature enough to detect profile faces reliably (and will be replaced for this reason in the near future). Frontal faces are precisely detected, which is mainly due to a good recognition of the eye region. This fact enables us to correct in-plane rotations of the head resulting from leaning the head to either shoulder (see Figure 4 for examples).

Figure 4, upper row: Examples of leant heads, leading to in-plane rotation of the face. The lower row shows the same faces after they have been rotated back to an upright position by *Videana's* face recognition system using the detected position of the eye region. This step is important for a later comparison of two faces.

After a first grouping phase, a further analysis is applied to persons who appeared in more than a predefined minimum number of shots. This analysis aims at finding face features that best discriminate this face group from the other groups. Finally, the classification is re-run using these group specific features. Preliminary results for the recognition of frontal faces are very promising. In particular, the correction of in-plane rotation and facial feature selection significantly improve the results: Based on a TV discussion sample, sufficiently large clusters could be built for 5 of the 6 appearing persons that could be used to represent a person and learn the specific facial features. The recall rate was 84% at 94% precision for the clustering (i.e. only 6% of the persons associated with a group did not conform to the group's main identity). The baseline system reached only a recall rate of 71% for the same precision score.

Applications of *Videana*

There are many applications to employ *Videana* for film studies. For example, we have conducted a case study in conjuction with the research project "Industrialization of Perception", which is also part of the Siegen research center. *Videana's* batch mode for shot segmentation allows users to analyze a number of videos automatically. This batch mode has been utilized to exemplarily analyze the cut frequencies of seven short films from the period of 1907-1913 from the USA and France. For example, it can be easily seen that the average cut frequency (ACF) of the American films is higher than that of the French films, and the ACF of the films of the 1911-1913 period is nearly twice as high as the ACF of the period 1907-1909. Of course, a larger number of videos is needed to obtain empirical evidence about such data. However, a tool such as *Videana* allows researchers to analyze a large number of videos with respect to

related research questions – annotating all these video manually would be a very time-consuming task.

Another project that applies *Videana* is "Media narrations and media games", also part of the Siegen research center. This project investigates hybrid forms of game and narration, which are observable in computer games and feature films since the 1990s. The aim is a formal-aesthetical and function-logical analysis of these sequences and a summarization into a typology. Besides supporting these research activities by means of the basic functionality provided by *Videana*, an extension is currently being developed which is able to learn certain sequence types or semantic concepts. For example, the underlying audio-visual data characteristics of narrative and playing sequences in computer games and feature films can be learned automatically. A possible application would be to let the software classify shots in such hybrid videos into narrative or interactive shots. In a next step, it could be analyzed which features allow to distinguish between these sequences and the others at the level of signal processing and machine learning. Of course, it would be left to media scientists to interpret these results.

Finally, the software toolkit *Videana* has been utilized for psychological research in an external cooperation, conducted together with Klaus Mathiak (RWTH University Aachen, Germany) and Rene Weber (University of California, Santa Barbara). An automatic semantic video analysis system was developed (Mühling et al 2007) to support interdisciplinary research efforts in the field of psychology and media science. The psychological research question studied is whether and how playing violent content in computer games may induce aggression. To investigate this question, the extraction of meaningful content from computer games is required to gain insights into the interrelationship of violent game events and the underlying neurophysiologic basis (brain activity) of a player. Previously, human annotators had to index game content according to the current game state, which is a very time-consuming task. The automatic annotation of a large number of computer game recordings (i.e. videos) speeds up the experimentation process and allows researchers to analyze more experimental data on an objective basis. The proposed computer game video content analysis system for computer games extracts several audiovisual low-level as well as mid-level features and deduces semantic content via a machine learning approach. This system requires manual annotations only for a single video to facilitate a semi-supervised learning process. Experimental results demonstrated the usefulness of the proposed approach for such research: 91% of the game events of inactive, preparation, search and violence could be recognized correctly.

Planned Extensions for *Videana*

There are several areas for future work. It is expected that exchanging the current face detector improves the results of face detection (and also recognition) of non-frontal faces. Furthermore, it is planned to combine face recognition with speaker recognition technology to exploit the audio stream of videos, in order to obtain a multimodal person recognition system. This way, the robustness of *Videana's* person indexing module will hopefully be improved.

On the other hand, we are further planning to investigate the detection of semantic concepts (for example, "indoor/outdoor" via exploitation of depth-information extracted from the video-stream, "war", "politician" etc.). This could be, together with the already implemented segmentation algorithms (shots, persons, audio), the basis of automatic storyline extraction, although this will probably not work in a fully automated manner in the near future.

Automatic Analysis versus Human Annotation

In this section, we briefly summarize the capabilities of state-of-the-art video analysis algorithms and compare the analysis performance of automatic computer systems with the quality of human annotations. In the field of shot boundary detection, state-of-the-art cut detection algorithms achieve recall and precision values of about 90% to 97%, whereas the detection of gradual transitions does not reach this quality and lies at approximately 80%. Similar results are achieved by camera motion estimation algorithms (80% to 90% recall at a precision of 95%). The research in the field of face detection is as good as described above, while recent algorithms demonstrate impressive recall and precision values of almost 100%. Up to date results in face recognition exhibit reasonable performance (90% recognition rate at a false alarm rate of 1%), whereas the recognition of persons in arbitrary video sequences is a clearly more challenging task. The performance of general semantic concept detection algorithms varies strongly depending on the type of the concept. The following examples show average precision values for some selected concepts, as they were obtained by the best systems at TRECVID's high-level feature detection task in 2005: map 53%, sports 52%, mountain 45%, car 37%, people walking/running 35%, US-flag 25%, explosion/fire 12%, and prisoner 5%.

Intuitively, one might think that humans always achieve a recognition rate of nearly 100%, but subjectivity and diminishing attention seem to be limiting factors. The comparison of manual annotations against each other shows the performance of automatic software systems in a more favourable light. Concerning some features, automatic analysis algorithms even outperform human

annotations. For example, the correlation of two manual annotations concerning camera motion revealed 66% recall at 100% precision for pan, respectively 34% recall at 94% precision for tilt (Bailer et al 2005), while automatic camera motion algorithms achieved 89% recall at 96% precision respectively 80% recall at 100% precision against another human annotation. Similar results can be observed in the field of shot boundary detection: in our experiments, the consensus of our (human) annotations lies between 80% and 97% with respect to the official (human) TRECVID annotation. Interestingly, these results are comparable to the best automatic systems evaluated at TRECVID.

A similar result can be observed in the psychological study of Weber et al (2006), which was later supplemented with our video analysis system. To be able to investigate interrelationships with the player's brain activity, the following game states of game sessions had to be annotated manually: inactive, preparation, search/exploration and violence. Weber et al (2006) report an inter-coder reliability of 0.85 for human annotators. Our automatic system demonstrates an excellent performance achieving an accuracy of up to 91% with respect to a human annotation.

Overall, it can be concluded that particular computer-based analysis approaches have reached a sufficient level of maturity and hence it is obvious that software tools can significantly aid scientific media analysis. In particular, they allow researchers to analyze larger data sets on an objective basis. Of course, humans still can do many things better, for example object/background separation, person recognition in arbitrary videos or generic object recognition, and last but not least, the qualitative interpretation of scene content will be reserved exclusively to humans for a conceivable period of time.

Acknowledgements

This work is financially supported by the Deutsche Forschungsgemeinschaft (SFB/FK 615, Project MT).

Bibliography

Amir, A., G. Iyengar, J. Argillander, M. Campbell, A. Haubold, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev, J. R. Smith, J. Tešić, and T. Volkmer. "IBM Research TRECVID-2005 Video Retrieval System." *TRECVID Online Proceedings* 2005. 31.08.2007. www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

Bailer, W., P. Schallauer, and G. Thallinger. "Joanneum Research at TRECVID 2005 – Camera Motion Detection." *TRECVID Online Proceedings* 2005. 31.08.2007. www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

Bescos, J. "Real Time Shot Change Detection Over Online MPEG-2 Video." *IEEE Transactions on Circuits and Systems for Video Technology* 1.4 (2004): 475-484.

Ewerth, R. and B. Freisleben. "Video Cut Detection without Thresholds." *Proceedings of the 11th International Workshop on Systems, Signals and Image Processing*. Poznan, Poland, 2004: 227-230.

Ewerth, R. and B. Freisleben. "Improving Cut Detection in MPEG Videos by GOP-Oriented Frame Difference Normalization." *Proceedings of the 17th International Conference on Pattern Recognition. Vol. 2*. Cambridge (UK) 2004: 807-810.

Ewerth, R., M. Mühling, and B. Freisleben. "Self-Supervised Learning of Face Appearances in TV Casts and Movies." *International Journal on Semantic Computing, Special Issue on ISM* (2006): 78-85.

Ewerth, R., M. Mühling, T. Stadelmann, E. Qeli, B. Agel, D. Seiler, and B. Freisleben. "University of Marburg at TRECVID 2006: Shot Boundary Detection and Rushes Task Results." *TRECVID Online Proceedings* 2006. 31.08.2007. www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

Ewerth, R., M. Schwalb, P. Tessmann, and B. Freisleben. "Estimation of Arbitrary Camera Motion in MPEG Videos." *Proceedings of the 17th International Conference on Pattern Recognition. Vol. 1*. Cambridge (UK) 2004: 512-515.

Giesenfeld, G., and P. Sanke. "Ein komfortabler Schreibstift für spezielle Aufgaben: Vorstellung des Filmprotokollierungssystems 'Filmprot' (Vers. 1.01)." *Filmanalyse interdisziplinär*. Ed. H. Korte, W. Faulstich. Second Edition. Göttingen: Vandenhoeck & Ruprecht, 1991: 135-146.

Gllavata, J., R. Ewerth, and B. Freisleben. "Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients." *Proceedings of 17th International Conference on Pattern Recognition. Vol. 1*. Cambridge (UK), 2004: 425-428.

Gllavata, J., R. Ewerth, and B. Freisleben. "Tracking Text in MPEG Videos." *Proceedings of ACM Multimedia*. New York, 2004: 240-243.

Gllavata, J., R. Ewerth, and B. Freisleben. "A Text Detection, Localization and Segmentation System for OCR in Images." *Proceedings of the 6th IEEE Int. Symposium on Multimedia Software Engineering*. Miami 2004: 310-317.

Gllavata, J., R. Ewerth, T. Stefi, and B. Freisleben. "Unsupervised Text Segmentation Using Color and Wavelet Features." *Lecture Notes on Computer Science: Proceedings of the 3rd International Conference on Image and Video Retrieval*. Dublin, 2004:216-224.

Hanjalic, A. "Shot Boundary Detection: Unraveled and Resolved?" *IEEE Transactions on Circuits and Systems for Video Technology* 12.2 (2002): 90-105.

OpenCV. Intel's Open Source Computer Vision Library. 31.08.2007. www.intel.com/technology/computing/opencv/

Jung, K., K. I. Kim, and A. K. Jain. "Text Information Extraction in Images and Video: A Survey." *Pattern Recognition* 37 (2004): 977-997.

Korte, H. "Projektbericht CNfA – Computergestützte Notation filmischer Abläufe – Erweiterte und aktualisierte Fassung." *IMF-Schriften* 1 (1992).

Korte, H. *Handbuch CNfA, Prototyp 3, Computergestützte Notation filmischer Abläufe*. Braunschweig: 1994.

Korte, H. "Einführung in die Systematische Filmanalyse." Berlin: Erich Schmidt, 2001.

Martinez, J. M. "MPEG–7 Overview." *Technical Report N4980, ISO/IEC JTC1/SC29/WG11*. Klagenfurt 2002.

Mühling, M., R. Ewerth, T. Stadelmann, B. Freisleben, R. Weber, and K. Mathiak. "Semantic Video Analysis for Psychological Research on Violence in Computer Games. *Proceedings of ACM International Conference on Image and Video Retrieval 2007 (CIVR 07)*. 2007: 611-618.

Phillips, P. J., P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. "FRVT 2002: Overview and Summary." 31.08.2007. www.frvt.org/FRVT2002/documents.htm.

Phillips, P. J., W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. "FRVT 2006 and ICE 2006 Large-Scale Results." March 2007. 31.08.2007. www.frvt.org/FRVT2006/Results.aspx.

Tahaghoghi, S. M. M., Thom, J. A., Williams, H. E., and Volkmer, T. "Video Cut Detection Using Frame Windows." *Proceedings of the Twenty-Eighth Australasian Computer Science Conference*. 38 (2005): 193-199.

Tsivian, Y., and G. Civjans. "CineMetrics.lv: Movie measurement and study tool database", 31.08.2007. www.cinemetrics.lv

TREC Video Retrieval Evaluation. 31.08.2007. www-nlpir.nist.gov/projects/trecvid/.

Truong, B. T., C. Dorai, and S. Venkatesh. "New Enhancements to Cut, Fade, and Dissolve Detection Processes in Video Segmentation." *Proceedings of the 8th ACM International Conference on Multimedia*. Marina del Rey, 2000: 219-227.

Vincent, L. "Announcing Tesseract OCR." *Google Code Blog, August 2006*. 31.08.2007. code.google.com/p/tesseract-ocr/.

Viola, P., and M. Jones. "Robust Real-Time Face Detection." *International Journal of Computer Vision*. 57.2 (2004): 137-154.

Weber, R., U. Ritterfeld, and K. Mathiak. "Does Playing Violent Video Games Induce Aggression? Empirical Evidence of a Functional Magnetic Resonance Imaging Study." *Media Psychology*. 8 (2006): 39-60.

Yang, M.-H., D. J. Kriegman, and N. Ahuja. "Detecting Faces in Images: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.1 (2002): 34-58.

Yeo, B., and Liu, B. "Rapid Scene Analysis on Compressed Video." *IEEE Transactions on Circuits and Systems for Video Technology* 5.6 (1995): 533-544.

Yuan, J., L. Xiao, D. Wang, D. Ding, Y. Zuo, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin, and B. Zhang. "Tsinghua University at TRECVID 2005." *Online Proceedings of TRECVID Conference Series 2005*. 31.08.2007. www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

Zhao, W., R. Chellappa, P. J. Phillips, and A. Rosenfeld. "Face Recognition: A Literature Survey." *ACM Computing Surveys* 35.4 (2003): 399-458.