THE MULTIMEDIAN CONCERT-VIDEO BROWSER

Ynze van Houten[1], *Umut Naci*[2], *Bauke Freiburg*[3],
Robbert Eggermont[2], *Sander Schuurman*[3], *Danny Hollander*[3], *Jaap Reitsma*[1], *Maurice Markslag*[1],
Justin Kniest[3], *Mettina Veenstra*[1] *& Alan Hanjalic*[2]

[1]Telematica Instituut
P.O.Box 589, 7500 AN Enschede, The Netherlands
[2]Delft University of Technology, Information and Communication Theory Group
Mekelweg 4, 2628 CD Delft, The Netherlands
[3]Stichting Fabchannel
Weteringschans 6-8, 1017 SG Amsterdam, The Netherlands

ABSTRACT

The MultimediaN concert-video browser demonstrates a video interaction environment for efficiently browsing video registrations of pop, rock and other music concerts. The exhibition displays the current state of the project for developing an advanced concert-video browser in 2007. Three demos are provided: 1) a high-level content analysis methodology for modeling the "experience" of the concert at its different stages, and for automatically detecting and identifying semantically coherent temporal segments in concert videos, 2) a general-purpose video editor that associates semantic descriptions with the video segments using both manual and automatic inputs, and a video browser that applies ideas from information foraging theory and demonstrates patch-based video browsing, 3) the Fabplayer, specifically designed for patch-based browsing of concert videos by a dedicated user-group, making use of the results of automatic concert-video segmentation.

1. INTRODUCTION

The real design issue regarding video interaction is greater efficiency in finding useful, interesting, or appealing video segments. The MultimediaN[1] concert-video browser demonstrates a video interaction environment for efficiently browsing video registrations of pop concerts as performed at the Dutch concert halls Paradiso and Melkweg, available at the Fabchannel website [1]. The exhibition displays the current state of the project for developing an advanced concert-video browser in 2007. This includes the development of an automatic video content analysis algorithm for providing non-linear access to semantically coherent video segments corresponding to their content and the "experience" they elicit, a multi-purpose patch-based video editing and browsing environment, and a mock-up of a concert-video browser with functionality and design serving a dedicated user-group.

Current interaction with concert videos – as in the old design of the Fabplayer - is mostly limited to selecting and playing a concert. This project aims at much more advanced user interaction, where users will be able to interact with smaller semantic units than complete concerts. These units are not limited to songs but also include smaller units with a particular affective (e.g., excitement) and cognitive (e.g. solo, vocals, instrumental) content. Detection of these semantically coherent temporal segments is preferably performed automatically. Next, the segments are semantically described, which may include attributes like who is performing, which instruments are played, what the performers look like (e.g. clothing), which texts are uttered etc. Thus, a set of video patches can be created. Video patches are collections of fragments sharing a certain attribute. For example, a video patch can be the collection of all fragments with guitar solos, or the collection of all fragments with stage divers, or the collection of all songs from a specific album, etc. The attributes are not randomly chosen, but will be acquired from end-users via surveys. Video fragments can come from one concert video, or from the whole collection of concert videos. The user can browse the layer of patches and have a rich interaction with the underlying concert-video database. The browsing activity itself can be the goal, or the user can search music fragments to create a playlist with favourite parts from one concert or from several concerts, which can then be replayed or shared with other users.

[1] MultimediaN is a Dutch national research program in the field of multimedia, 2004-2009

0-7803-9332-5/05/$20.00 ©2005 IEEE

2. RELATION TO SCIENTIFIC RESEARCH

The development of the video browser mainly relies on research on efficient video interaction. We apply ideas from human-information interaction theories - most notably information foraging theory [2]. According to this theory, people perceive their information environment as being "patchy" (compare websites on the World Wide Web). People can stay within a patch, or switch to another patch. Browsing patches increases efficient interaction as other video content can be (temporarily) ignored. Links to other patches ("browsing cues") are constantly provided, facilitating users to switch to other patches or to combine patches. When a browsing cue has a semantic match with a user's goals or interests, this cue carries a "scent" for that user. In this theory, it is stated that people browse the information environment – in this case, video material - by following scent. A first prototype applying these ideas is described in [3]. The current video browsing application builds on this research, and is used within the Dutch MultimediaN program to study browsing behaviour.

The mechanism enabling automatic creation and identification of patches is based on a novel methodology for high-level temporal content analysis of a concert video. The scientific challenge and relevance of developing automated content analysis solutions for the specific application domain of music concert video are large due to the highly rich and complex content to be analyzed. This content is characterized by "fuzzy" boundaries between temporal semantic segments, large diversity of semantic segment categories at different granularity levels (e.g. "song", "instrumental solo", "vocal solo", "group singing", "vocals and instruments", etc.) and high noisiness of the semantic segments, which makes their proper detection and classification very difficult. In other words, the music concert video can be seen as an excellent test bed for pushing the state-of-the-art techniques for high-level video parsing and classification beyond their current limits.

3. DEMO DESCRIPTIONS

3.1. The concert-video parsing algorithm

The processing of a concert video for enabling non-linear content access starts with applying a newly developed methodology for 1) modeling the "experience" of the concert, 2) automatically detecting boundaries of consecutive, semantically coherent temporal segments (the items of a patch) at different levels of hierarchy, and 3) classifying detected patches as "instrumental solo", "vocal solo", "group singing", "vocals and instruments", etc.

We approach the modeling of the "experience" of a concert by extending our previous work on arousal modeling [4]. Based on a number of audio-visual and editing features, the effect of which on a human viewer can be related to how that viewer "experiences" different parts of the concert, we model the arousal time curve that represents the variations in experience from one time stamp to another. High arousal values ideally represent the parts of a concert with high excitement, as compared to more-or-less "serene" songs or song parts represented by low arousal values. The obtained curve can be used to automatically extract the parts of the concert that are best capable of eliciting a particular "experience" in the given total duration.

Patches are created and identified based on a combined algorithm for high-level temporal parsing and classification of concert video material. This algorithm aims at detecting time stamps, which can be seen as logical boundaries separating semantic temporal units of a video at different levels of hierarchy. At the top of the hierarchy, boundaries separating the songs are found. At lower levels, boundaries of smaller semantic units, such as, "instrumental solo", "instrumental section" or "purely vocal" sections, are detected. We approach the development of the combined parsing and classification framework using the features from both the audio and visual modality. For performing audio content analysis, we build on extensive previous work, good surveys of which can be found in [5], [6] and [7]. The oldest methods proposed in this field aimed at speech-music discrimination. Sounders [8] reported the performance of 98% by using only zero crossing rate (ZCR) and short-term energy as features, and by working with hard-set thresholds. Zhang and Kuo [9] also used ZCR and energy function together with fundamental frequency and spectral peak information for detecting speech, song, music, silence and background noise, by using a heuristic rule-based system. The methods proposed by Moreno and Rifkin [10] and Seck et al. [11] are examples of systems using cepstrum-based features. Lately the research [12], [13] and [14] has started to focus on detection of special audio effects like applauses, gunshots, car-crashes, helicopter sound etc. Although rather high quality of performance was reported in many of the abovementioned papers, we found that the specific application of concert video analysis required adaptation of the existing techniques to be able to deal in a robust way with the extremely rich and complex (noisy) concert audio content. The selection of visual features, on the other hand, primarily aims at utilizing the domain knowledge about the lighting conditions during the concert and expected camera operations at various stages of a concert. For instance, it is rather usual that the light show follows the rhythm of the music. Further, one could expect a close-up of a musician performing a solo. Finally, the break between songs is often accompanied by a wide-range camera shot, and by less dynamic or even stationary light show. The audio and visual data streams of a concert video are processed separately and the result of video processing are used to correct the boundaries and classification results obtained by audio analysis.

3.2. The general-purpose video editor and browser

The general-purpose video editor is used to enrich the video with metadata. The format chosen for the metadata is MPEG-7 [15]. Apart from editing the creation and media information of the video, the most important capability of the editor is the ability to define a segmentation hierarchy. At the lowest level the output of parsing algorithms is used, e.g., video shot detection, instrument detection and more. The low level segments are combined at a higher level in semantic segments, e.g., a song. All segments can be annotated with a description in text. Additionally, semantic attributes can be attached to a segment. The attributes are references to a predefined ontology. For easy visual identification of segments, support will be added to assign one or more key frames to a segment.

The video editor is developed in Java as a collection of plugins for the Eclipse platform [16]. The prototype uses the Java Media Framework for playing the video and for the extraction of frames, but an implementation using an embedded Microsoft Media Player for playing the videos is also available. Although many controls are available in SWT (the Standard Widget Toolkit for Eclipse), no control is available yet for operating on time-based media. A custom control was developed to be able to view and edit video segments along a timeline, including the ability to zoom in and out. The graphical part is connected to the underlying MPEG-7 data by the Eclipse Modelling Framework (EMF). EMF is a modeling framework and code generation facility for building tools and other applications based on a structured data model. The inputs for the EMF are the MPEG-7 XSD schemas. EMF generates the Java interfaces and implementations for accessing the MPEG-7 elements. EMF also includes a command framework supporting undo/redo functionality.

The editor is essentially file-based, using the open source XML database eXist [17] as the repository. The editor is planned to access the database through the WebDAV protocol for checking in and out the metadata files.

The video browser is also a client application built on top of the Eclipse platform. In comparison with the editor the look will be more like a traditional browser. The browser shares most of its plugins with the editor.

The video browser allows a user to search and browse through a video collection. Search can be text-based or patch-based. A text-based search will use the segment annotations, whereas the patch-based search uses the semantic attributes attached to the segments. The prototype will use the search capabilities of the eXists XML database. The queries are formulated in XQuery. The full-text search uses a database-specific extension.

Results from a query or a patch selection are displayed in various ways. Videos or video segments in the result set are represented in different ways by video abstractions, including texts (title, descriptions) and pictures (keyframes, storyboards). Results are displayed as a list or on a timeline. One specific video segment can be viewed in more detail, where users can switch between different display modes. Links to video segments are displayed in such a way that users can easily judge the relevance of the related video segment (that is, whether the link carries “scent”). An important aspect of the browser is that when inspecting the fragments in a patch (“within-patch browsing”), of each segment that is inspected it is displayed which other patches it also belongs to. This way, users are encouraged to switch from patch to patch (“between-patches browsing”).

3.3. The Fabplayer

Fabchannel [1] broadcasts concerts, festivals, competitions and lectures from Paradiso and Melkweg (Amsterdam) on the internet. Besides live broadcasts Fabchannel offers an extensive online video archive including over 330 recordings. This makes Fabchannel one of the biggest streaming video initiatives in the world regarding live music.

Besides broadcasting many big international acts - like the Stereophonics, Presidents of the USA and Damien Rice - Fabchannel is always on the lookout for new, less known, but talented bands that are on the edge of breaking through. By broadcasting these bands and spreading their music Fabchannel helps them to reach a broader audience. For this reason Fabchannel’s streaming video archive should be very easily accessible. The Fabplayer project is working towards new ways of exploring the archive.

The Fabplayer is the video player framework developed by the research and development department of Fabchannel. Fabchannel is currently designing a new version of the Fabplayer that will support intelligent video browsing within recorded concerts. The Fabplayer introduces new ways of navigating concert videos, for example by song, solos, and intensity. The demo of the Fabplayer presented at the ICME 2005 exhibition will be the first draft of a working interaction design. The demo will demonstrate how the Fabchannel audience can interact with the concert videos in an entertaining way.

Fabchannel users can not only find and watch (parts of the) concerts in the Fabplayer, they are also encouraged to enrich the video by annotating the available 'units' or 'patches' within the video. For example, users can attach songtitles to songs or describe which artist is playing a solo on what instrument. This kind of user participation should lead to more detailed searching, browsing, and navigating options in upcoming versions of the Fabplayer.

Fabchannel attracts music fans from all over the world. Often they are part of online communities like artists' fansites. These fans know a lot about the artists and their performances. With the new Fabplayer this information can be captured and stored as metadata along the timeline of a concert video.

Together with automated video content analysis and professional annotation using the general-purpose video editor, this kind of user participation will lead to interesting ways of experiencing streaming video concerts.

4. CONCLUSION

This paper describes the early stage of the development of a concert-video browser which makes use of automatic concert-video segmentation and which supports patch-based video browsing. The three demos span the whole dimension from generic scientific research to a specific application serving a specific user group. The general-purpose video browser and the Fabplayer will be used in future studies of video browsing behaviour.

5. REFERENCES

[1] http://www.fabchannel.com

[2] Pirolli, P. & Card, S. K. (1999). Information foraging. *Psychological Review, 106*, 643-675.

[3] van Houten, Y., Schuurman, J. G., & Verhagen, P. (2004). Video content foraging. In P. Enser, Y. Kompatsiaris, N. O'Connor, A. F. Smeaton, & A. W. M. Smeulders (Eds.), *Lecture Notes in Computer Science 3115 - Image and Video Retrieval - Proceedings of CIVR 2004* (pp. 15-23). Springer-Verlag.

[4] Hanjalic A., L.-Q. Xu: Affective video content representation and modeling, IEEE Transactions on Multimedia, Vol.7, No.1, pp.143-154, February 2005

[5] Wang Y., Z.Liu and J.C. Huang, "Multimedia content analysis: Using audio and visual clues", *in IEEE Signal Processing Magazine*, vol 17, no. 6, pp. 12-36, Nov. 2000.

[6] Zhang, T., and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *in IEEE Transactions on Speech and Audio Processing*, Vol 9. No 4, 2001.

[7] Snoek C.G.M. and M. Worring, "Multimodal video indexing: A review of the state-of-the-art", *in Journal of Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5-35(31), January 2005.

[8] Saunders J., "Real-time discrimination of broadcast speech/music," in Proc. IEEE ICASSP, 1996.

[9] Zhang T. and J. Kuo, "Audio content analysis for on-line audiovisual data segmentation and classification,", *in IEEE Trans. Speech Audio Processing.*, vol. 9, no. 3, pp. 441–457, May 2001.

[10] Moreno P. and R. Rifkin, "Using the fisher kernel method for web audio classification," *in Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1921–1924, 2000.

[11] Seck M., F. Bimbot, D. Zugah, and B. Delyon, "Two-class signal segmentation for speech/music detection in audio tracks," *in Proc. Eurospeech*, pp. 2801–2804, September, 1999.

[12] Moncrieff S., C. Dorai and S. Venkatesh, "Detecting indexical signs in film audio for scene interpretation", in Proc. IEEE ICME'01, pp. 1192-1195, Tokyo, Japan, August, 2001.

[13] Baillie M., Jose J.M., "An audio-based sports video segmentation and event detection algorithm", Event Mining 2004: CVPR 2004 Workshop, IEEE Computer Society, Washington DC, July, 2004.

[14] Cai R., L. Lu, A. Hanjalic, H.-J. Zhang, L.-H. Cai, "A flexible framework for key audio effects detection and auditory context inference", *IEEE Transactions on Speech and Audio Processing*, 2005.

[15]http://www.medialab.sonera.fi/workspace/MPEG7WhitePaper.pdf , MPEG-7 White Paper, Sonera MediaLab, October 13, 2003.

[16] http://www.eclipse.org

[17] http://www.exist-db.org