JPEG2000 BASED SCALABLE SUMMARY FOR REMOTE VIDEO CONTENT BROWSING AND EFFICIENT SEMANTIC STRUCTURE UNDERSTANDING[c]

J. MEESSEN

Multitel asbl
Avenue Copernic,1
7000 Mons, Belgium
E-mail: Jerome.meessen@multitel.be

L.-Q. XU

Broadband Applications Research Centre
BT Research & Venturing
Adastral Park, Ipswich IP5 3RE, UK
E-mail: li-qun.xu@bt.com

B. MACQ

Communication and Remote Sensing Lab.
UCL
Louvain-la-Neuve, Belgium
E-mail: macq@tele.ucl.ac.be

This paper presents a new method for remote and interactive browsing of long video sequences. The solution is based on interactive navigation in a scalable mega image resulting from a JPEG2000 coded keyframe-based video summary. The presented system is compliant with the new JPEG2000 part 9 "JPIP – JPEG2000 Interactivity, API and Protocol," which lends itself to working under varying channel conditions such as wireless networks. The flexibility offered by JPEG2000 allows the application to highlight interactively keyframes corresponding to the desired content first within a low quality and low-resolution version of the full video summary. It then offers a fine grain scalability for a user to navigate and zoom in to particular scenes or events represented by the keyframes. This possibility to visualise keyframes of interest and playback the corresponding video shots within the context of the whole sequence enables the user to understand the temporal relations between semantically similar events, i.e. a new way to analyse long video sequences.

[c] This work was partially supported by the EU FP5 project SCHEMA, IST-2001-32795.

1. Introduction

With the advent of digital revolution, the increasing network connectivity and bandwidth, as well as the mass production and distribution of rich audio-visual media, the demands from end users are becoming ever more urgent for a fast and easy access to video program summaries in order to browse and visualize desirable contents [1]. A video content summary often takes the form of a 2-D presentation on a visualisation interface, which is made up of selected frames, or keyframes, representing semantically related data chunks, i.e. shots or events. Depending on the media genre and applications, many different layouts for keyframes presentation are possible, as discussed in [2].

However, summarizing the content of a long video sequence this way for entertainment genres such as a feature movie or drama, the number of selected keyframes is still way too many. Two problems are ensued. First, the semantic story structures will be largely buried in the numerous images displayed. And, secondly, in the case of a user accessing the summary stored on a remote server, the transmission over the network of a large number of keyframes is a major problem. Today, there exist different approaches to addressing these issues. Building condensed and semantically relevant video summaries has been seen in the work by Yeung and Yeo [3] and by Chiu et al. [4]. However, though these summaries provide a good overview of a sequence, they tend to present to the user one pre-defined semantic level illustration of the sequence only. Hierarchical shots clustering and presentation is another common approach to browsing either one video sequence [5][6][7] or a video sequences database [8]. In particular, the shot clustering and browsing methods described [9] and [10] are particularly interesting since they are evaluated regarding the amount of transmitted data at each user retrieval request. . These are efficient solutions to getting a quick overview of a video sequence and to finding a particular scene of interest. However, they do not have provisions for contextual visualization of the links between semantically similar scenes, i.e., to answer a user's queries like "What happened before and after that particular event?" "Are there any other similar events taking place in the story, and if so, what are their temporal relations?" etc.

In this paper, we focus on helping the user to understand the underlying semantic structure of a video sequence, i.e. to establish relations between semantically similar scenes and highlight them within the context of the whole sequence. Rather than to propose a complex shot clustering strategy or storyboard layout, we exploit the user's intelligence by providing him/her with interactive tools for intuitive navigation in a remotely stored scalable summary. The idea is to exploit the powerful features of compression and scalable representation in JPEG2000, the new standard for still image compression [11], and produce scalable keyframe-based summaries of a video sequence, while allowing for at the same time semantics-based queries. JPEG2000 allows a

flexible access to each spatial region of the compressed image, at a different resolution and PSNR quality level [12]. This is particularly suited to browsing very large images as discussed in [13]. We present here a layered platform compliant with the forthcoming JPEG2000 Part 9 "JPIP – JPEG2000 interactive protocol" [14][15]. While storing only one detailed keyframe based video summary, or storyboard, this standard communication between server and client provides the means to transmit efficiently many different versions of the storyboard over a network. Moreover, JPIP offers means to adapt the transmission to changing channel conditions allowing an efficient transmission of the summary data with any type of channel conditions and user processing resources. This particularly suits video browsing using mobile devices.

The annotation of shots and scenes of the video summary is based on MPEG-7 visual content description schemas [16]. After the temporal decomposition of a video sequence into segments, i.e. shots or scenes, and necessary manual annotation of their contents, an MPEG-7 compliant XML description file specifies, for each of these segments, a number of attributes, including the text annotations (scene, object, action) and time information – the start and duration of the segment, and the position of the keyframe selected for the shot [17]. The annotations of shots allow translating content-based queries into image-oriented requests.

2. System Framework

This section discusses the two core components underlying the proposed video content browsing and retrieval system.

2.1. *Scalable keyframe-based video summary with annotation*

The work flow used to create the coded keyframe-based video summary is as follows. The original video sequence is segmented into shots, and one keyframe is selected for each shot to represent its visual content. The representation scheme can be extended to a group of shots, or sub-scene or scene, to avoid the redundancy in displayed visual content, as in the case of "A Beautiful Mind" video discussed in the experimental section. The keyframe images are then arranged in raster scanning order to compose a large mosaic image, which is then JPEG2000 compressed to output two files: the JPEG2000 codestream and its associated index file as defined in JPEG2000 Part 9 'JPIP'. The compression is done with at least two quality levels, to be able to highlight keyframes of interest, and with different resolution levels. Moreover, the JPEG2000 coding parameters are chosen such that each keyframe can be accessed separately.

Automatic detection of shot boundaries and the selection of keyframe(s) for each shot / scene are carried out using the segmentation and annotation tool described in [18]. In case of errors, the shot segmentation results can be

manually edited by splitting and merging shots. The meanings of each shot are annotated using a set of keywords from a predefined hierarchical lexicon. The annotations are saved in an MPEG-7 compliant XML file.

2.2. *System architecture*

Figure 1 presents the proposed client-server system architecture, which is based on the IST PRIAM project [19].

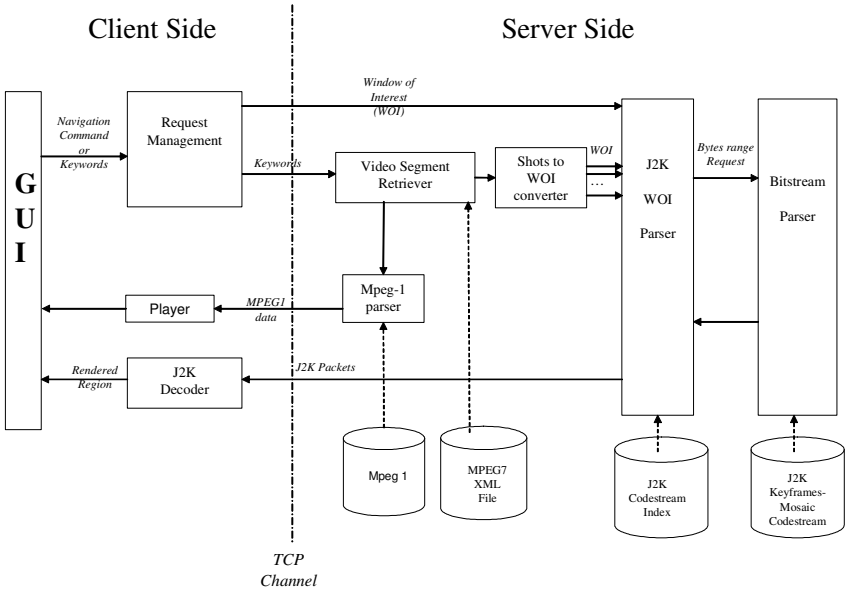

Figure 1. Client-server system architecture

We consider two types of client requests: the navigation request and the retrieval request.

The navigation requests (zooming, panning etc) are translated at the client side into a request for Windows of Interest (WOI) [13]. A WOI specifies a spatial region, a quality level and a resolution level. At the server side, the JPEG2000 WOI parser converts this WOI request into the selection of relevant JPEG2000 packets using the codestream index file. These packets contain additional data improving the quality of the requested regions once transmitted and decoded at the user's side.

The retrieval queries, based on the keywords from a predefined lexicon, are linked at the server side with shots indices by searching through the MPEG-7 annotation file of the video summary. The selected shots are then associated to WOI's. The WOIs' spatial region is defined by the shot's keyframe size and position in the mosaic. The corresponding WOIs specify the highest available

quality and resolution levels so as to highlight keyframe of interest as much as possible.

A video player module is also implemented to allow the user to play a particular shot/scene of interest, after browsing through the highlighted images.

3. EXPERIMENTS

In this section we discuss the application scenarios of the proposed system and present the experimental results obtained.

3.1. *Typical scenario*

A typical application scenario is described as follows. First, the GUI displays an overview of the JPEG2000 coded mosaic, which is obtained by decoding only a greyscale version of the lowest resolution and quality levels. The user then selects the keywords from the annotation dictionary and requests the server to present certain desired and more detailed contents of the video. The relevant keyframes, retrieved by the server will then be highlighted by the GUI. Enhancing the visual quality and resolution of keyframes within the initial low quality overview clearly shows the temporal semantic links among the contents of these shots and scenes. The user can also pan and zoom in the video summary and choose to play the video clip of a particular shot of interest.

3.2. *Preliminary trials*

To evaluate the functionality of the prototype system, experiments are performed on two movie excerpts; one from *"Notting Hill"* and the other from *"A Beautiful Mind"*. Table 1 specifies the attributes of the two video sequences including their respective keyframe-based video summary as well as the compressed summary size. A snapshot of the system in action is shown in Figure 2.

Table 1. Characteristics of the chosen video excerpts and the associated keyframe-based summary

Feature movies	"Notting Hill"	"A Beautiful Mind"
Time length	20 min	40 min
Keyframes size	352×288	352×240
No. of keyframes	16×15	13×13
Summary dimensions	5632×4320	4576×3120
Summary size	191 MB	192 MB
Compressed summary size	1400 KB	837 KB

Figure 2 Screen shots of the system on the video excerpt from the movie “A beautiful Mind”. The keyframes corresponding to the retrieval query "Urban" are highlighted within the overview (top). Thanks to JPEG2000, low cost panning and zooming in the summary are permitted. Scenes of interest can be visualised using the video player module (bottom) by clicking on the highlighted frame regions, respectively.

4. CONCLUSIONS

A new method for building a scalable representation of keyframe-based storyboard is proposed by exploiting the powerful compression and scalability of JPEG2000. Moreover, we have extended a JPEG2000 Part 9 (JPEG2000 Interactive Protocol) compliant platform to browse interactively the video summary, which would be accessible under different networking conditions or

processing resources. Using the MPEG-7 description scheme for annotating the semantic content of the sequences, the proposed system allows the user to visualize the links between semantically similar scenes, i.e. a new way to understand long video sequences. The prototype system has been tested using two long movie excerpts.

The approach exploits the user's interpretation capability while keeping the video summarization and description very simple.

Further work to extend the existing features is necessary, including the provision of functionalities allowing for real-time streaming video at the client side, enabling queries by example, search for desired content, and still more advanced semantic search mechanism. Moreover, this study will be used in the challenging context of video surveillance where the relation between similar scenes is a critical issue.

References

[1] A. Smeaton, "Challenges for the Content-Based Navigation of Digital Video in the Fishlar Digital Library," *Proc. of CIVR'2002*, London, July 2002.

[2] H. Lee, A. Smeaton et al: "Implementation and Analysis of Several Keyframes-Based Browsing Interfaces to Digital Video," *Proc. of the 4th European Conference on Digital Libraries (ECDL)*, Lisbon, Portugal, pp. 206-218, Sept. 2000.

[3] M. Yeung and B. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," *IEEE Trans. On Circuits and Systems for Video Technology*, **7**(5), October 1997.

[4] P. Chiu, A. Girgensohn and Q. Liu, "Stained-Glass Visualization for Highly Condensed Video Summaries," *Proc. of IEEE International Conference on Multimedia and Expo* (ICME'04), Taipei, Taiwan, June 2004.

[5] A.M. Ferman and A. M. Tekalp, "Two-Stage Hierarchical Video Summary Extraction to Match Low-Level User Browsing Preferences," *IEEE Trans. on Multimedia*, **5**(2), pp 244-256, June 2003.

[6] F. Shipman, A. Girgensohn and L. Wilcox, "Generation of Interactive Multi-level Video Summaries," *Proc. of ACM Multimedia 2003*, Berkeley, USA, Nov. 2003.

[7] J. Fan, A. Elmagarmid, X. Zhu, W. Aref and L. Wu, "ClassView: Hierarchical Video Shot Classification, Indexing, and Accessing," *IEEE Trans. on Multimedia*, **6**(1), pp. 70-86, Feb. 2004.

[8] C. Taskiran, J-Y Chen, A. Albiol, L. Torres, C. Bouman and E. Delp, "Vibe: A Compressed Video Database Structured for Active Browsing and Search," *IEEE Trans. on Multimedia*, **6**(1), pp. 103-118, Feb. 2004.

[9] J. R. Smith, "VideoZoom Spatio-temporal Video Browser," *IEEE Trans. on Multimedia*, **1**(2), June 1999.

[10] A. Doulamis and N. Doulamis, "Optimal Content-Based video Decomposition for Interactive Video Navigation," *IEEE Trans. on Circuits and Systems for Video Technology*, **14**(6), pp. 757-775, June 2004.

[11] ISO/IEC 15444-1 JPEG 2000 image coding system Part 1: Core coding system.

[12] D. Taubman and M. Marcellin, "JPEG2000: Standard for interactive imaging," *Proceedings of the IEEE*, **90**(8), pp. 1336-1357, Aug. 2002.

[13] J. Meessen, T. Suenaga, M. Iregui Guerrero, C. De Vleeschouwer and B. Macq, "Layered architecture for navigation in JPEG2000 Mega-Images," *Proc. of the 4th European Workshop in Image Analysis for Multimedia Interactive Services* (WIAMIS'03), pp. 92-95, London, UK, April 2003.

[14] JPIP Editors, "JPEG 2000 image coding system – Part 9: Interactivity tools, APIs and protocols – Final Committee Draft 2.0," *http://www.jpeg.org /public/fcd15444-9v2.doc*

[15] D. Taubman and R. Prandolini, "Architecture, philosophy and performance of JPIP: internet protocol standard for JPEG2000," *Proceedings of the International Symposium on Visual Communications and Image Processing* (VCIP'2003)*, 2003.*

[16] T. Sikora, "The MPEG-7 Visual Standard for Content Description - An Overview," *IEEE Trans. on Circuits and Systems for Video Technology*, **11**(6), pp. 696-702, June 2001.

[17] P. Salembier and J. Smith: "MPEG-7 multimedia description schemes," *IEEE Trans. on Circuits and Systems for Video Technology*, **11**(6), pp. 748-759, June 2001.

[18] IBM Research, "VideoAnnEx, Annotation Tool": available online at URL: *http://www.research.ibm.com/VideoAnnEx/.*

[19] EU IST FP5 project PRIAM (IST28646) "Platform for Real-Time and Interactive Access to Mega-images," *http://www.tele.ucl.ac.be/PROJECTS /PRIAM*.