

# JPEG 2000 Based Scalable Summary for Understanding Long Video Surveillance Sequences

Jerome Meessen<sup>a</sup>, Jean-Francois Delaigle<sup>a</sup>, Li-Qun Xu<sup>b</sup> and Benoit Macq<sup>c</sup>.

<sup>a</sup> Multitel asbl, 7000 Mons, Belgium  
{jerome.meessen, delaigle}@multitel.be

<sup>b</sup> BT Research & Venturing, Adastral Park, Ipswich, UK  
li-qun.xu@bt.com

<sup>c</sup> Communication and remote sensing lab., Univ. catholique de Louvain, 1348 Louvain-la-Neuve, Belgium  
macq@tele.ucl.ac.be

## ABSTRACT

This paper presents a new method for remote and interactive browsing of long video surveillance sequences. The solution is based on interactive navigation in JPEG 2000 coded mega-images. We assume that the video ‘key-frames’ are available through automatic detection of scene changes or abnormal behaviors. These key-frames are concatenated in raster scanning order forming a very large 2D image, which is then compressed with JPEG 2000 to produce a scalable video summary of the sequence. We then exploit a mega image navigation platform, designed in full compliance with JPEG 2000 part 9 “JPIP”, to search and visualize desirable content, based on client requests. The flexibility offered by JPEG 2000 allows highlighting key-frames corresponding to the required content within a low quality and low-resolution version of the whole summary. Such a fine grain scalability is a unique feature of our proposed JPEG 2000 video summaries expansion. This possibility to visualize key-frames of interests 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. It is then particularly suited to analyzing complex incidents consisting of many successive events spread over a long period.

Keywords: Video browsing, contextual understanding, scalable summary, video surveillance, JPEG 2000, JPIP.

## 1. INTRODUCTION

Nowadays the number of surveillance applications relying on digital video processing increases drastically. Consequently, the demand from users is becoming ever more urgent for a fast and easy access to video events summaries in order to browse and visualize desirable contents stored within archived databases. This paper addresses this issue by focusing on understanding complex events, i.e., several similar events spread over a long recorded period.

A video content summary often takes the form of a 2-D presentation on a visualization interface, which is made up of selected frames, or key-frames, representing semantically related data chunks, i.e. shots, events or scene changes. Fortunately, the key-frame selection in video surveillance can be achieved automatically using scene analysis techniques, see e.g. [1]. Given that the key-frames are available, depending on the application scenarios, many different layouts for key-frames presentations are possible, as discussed in [2]. However, summarizing the content of a long video surveillance sequence in this way appears to have the weakness that the temporal relations between similar semantic events are lost. This may be a critical problem for the end user trying to understand the contextual information of the

events occurred. Moreover, in the case of a user browsing a remotely stored video sequence, the amount of transmitted data can be way too expensive in terms of transmission costs.

Today, there exist different approaches to addressing these issues in the context of browsing structured video, like entertainment movies or broadcast archives. Building condensed and semantically relevant video summaries has been seen in the work by Yeung and Yeo [3] and by Chiu et al. [4]. Although these summaries provide a good overview of a sequence, they tend to present to the user only one pre-defined semantic level illustration of the sequence. In the case of automatic scene analysis for surveillance video, this is often not sufficient. Indeed, the resulting scenes understanding and annotations do not correspond to the semantic interpretation level of the user's, which is often much higher. Consequently, the video summary will not be useful.

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 in [9] and [10] are very interesting since they are evaluated regarding the amount of transmitted data at each user retrieval request, which is one of the problems we are concerned with. 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 recorded period, and if so, what are their temporal relations?" etc. One can imagine how important this contextual understanding can be in surveillance applications. Once the content of interest is retrieved, there is often the need to visualizing the preceding and following events, i.e. temporal context, as well as finding similar events that occurred earlier or later in the video sequence under scrutiny, i.e. the semantic context. As an example, let us consider a video sequence acquired in a train station and the event of interest is 'abandoned luggage' that can be detectable. This event will be attached higher degree of importance when only one 'abandoned luggage' is detected over the sequence, i.e., the event is abnormal, as compared with the case that there are 15 pieces of 'abandoned luggage' being detected within the same sequence, i.e., the event is fairly normal. Similarly, a 'fight' detection has an increased importance if other 'fights' are observed in a given temporal neighborhood. As a consequence, an efficient method for retrieving one event, as provided by hierarchical browsing, is not relevant enough compared to presenting the events within a contextual overview of the sequence.

In this paper, we focus on helping the user to understand the semantic and temporal context of events, i.e. to enable visualization of relations between semantically similar scenes changes or events and highlight them within the context of the whole sequence. Rather than 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. In doing this, we avoid freezing the video summary at a non-satisfactory level of semantics while minimizing the amount of visual data to be transmitted to the user.

The idea is to exploit the powerful features of compression and scalable representation in JPEG 2000, the new standard for still image compression [11], and produce scalable key-frame-based summaries of a video sequence, while allowing for at the same time semantics-based queries. JPEG 2000 offers a highly scalable representation of the compressed image, in terms of image components, spatial access, resolution and PSNR quality [12]. This is particularly suited to browsing very large images as discussed in [13] and [16]. The resolution scalability is due to a Discrete Wavelet Transform (DWT). The PSNR quality scalability is done by the optimal codestream decomposition in different quality layers[14]. The spatial access to the image can be obtained thanks to different possible mechanisms. The source image can be tiled in smaller images, i.e. tiles, that are compressed independently and whose contributions are signaled in the codestream. Another way to access spatial regions is provided by the precinct partitioning of the DWT coefficients [15]. Within the codestream, a packet contains the data corresponding to one precinct at a given resolution level and from one quality layer.

Here, we present a layered platform compliant with the new JPEG 2000 Part 9 "JPIP – JPEG 2000 interactive protocol" [17][18]. While storing only one detailed key-frame based video summary, or storyboard, this standardized communication between server and client provides the tools for exploiting the JPEG 2000 flexibility, and accessing interactively many different versions of the storyboard over a network. JPIP uses a codestream index, which basically details the coding options used for creating the codestream, i.e. number of quality layers, resolution levels etc., and contains the position, in bytes, of the JPEG 2000 codestream elements like headers, precincts, tiles etc. More details

about the structure of the codestream index file can be found in Annex I of the JPEG 2000 Part 9 standard [17]. 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.

Though a video shot, i.e. a continuous camera recording, is normally considered as the data unit to be represented by a key-frame in the case of post-production video program summary and browsing, this can hardly be true in video surveillance and monitoring situations. In most cases indeed, a surveillance sequence is made up of only ‘one’ single camera shot, continuously recorded by a fixed camera. To address this issue, Kim and Hwang proposed in [1] to base surveillance sequences abstraction on automatic object detection. In this work, we choose to associate the key-frame selection with each scene change, i.e. each modification of the dynamic scene content. For instance, one person walking in one direction and, later, changing his or her direction will correspond to 2 key-frames.

The annotation of scenes of our video summary is based on MPEG-7 visual content description schemas [19]. After the temporal decomposition of a video sequence into segments, i.e. scenes, and necessary manual or automatic 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 key-frame selected for the shot [20]. The annotations of shots allow translating content-based queries into image-oriented requests as detailed in section 2.2. It is worth noting that the JPEG standardization group has recently launched a new work item, named JPSEARCH, whose goal is to extend the JPEG 2000 XML metadata fields in order to enable new content-based applications like image retrieval [17]. Since our annotations structure is based on XML schemes, it could easily be modified to be compliant with this upcoming standard.

The paper is organized as follows. The next section details both the proposed method for building a scalable summary and the browsing system architecture. Section 3 describes the conducted experiments on some publicly available video sequences. The paper concludes in Section 4 with a discussion of future research work to enhance the system.

## **2. SYSTEM FRAMEWORK**

This section discusses the two core components underlying the proposed video content browsing and retrieval system. Firstly, the proposed method for creating and annotating a scalable key-frame based video summary is presented (section 2.1) and secondly, the system architecture allowing content retrieval and interactive browsing of the summary is described (section 2.2).

### **2.1. Scalable key-frame-based video summary with annotation**

The workflow used to create the coded key-frame-based video summary is depicted in Figure 1. The original video sequence is segmented into scenes, and one key-frame is selected for each scene to represent its visual content.

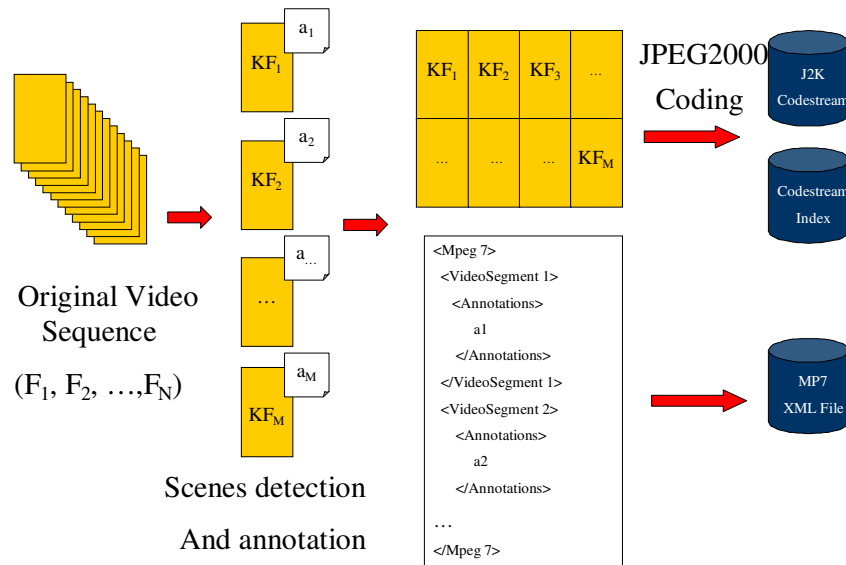


Figure 1. The creation of a scalable key-frame-based annotated video summary, which is coded by JPEG2000 in preparation for interactive system access

The decomposition into scenes can be done automatically using video analysis techniques. The proposed creation of scalable summary can also be applied on top of other key-frame selection methods. However, our goal is to provide a global representation of the content, helping the user to get a contextual understanding of events of interest. A too coarse content abstraction would result in an inefficient summary. We then suggest using a fine granularity for the scene change detection. In the case of surveillance video where most of the time no abrupt scene change is observed, this can be a major issue that can be solved by manual refinement or specific tuning of the video analysis process. Even if this problem is common to all automatic video analysis application, we consider it as the major constraint for our proposed system.

Once the key-frames images are selected, they are arranged in raster scanning order to compose a large mosaic image, which is then JPEG 2000 compressed to output the JPEG 2000 codestream, i.e. the coded data stream, and its associated index file as defined in JPEG 2000 Part 9 'JPIP'. The summary compression is done with at least two quality layers, and different resolutions levels, to be able to highlight, i.e. to improve the visual quality, key-frames of interest compared to the other key-frames. Moreover, the JPEG 2000 tiles and precincts dimensions are chosen such that each key-frame can be accessed separately. Since the JPEG 2000 coding standard imposes the precincts' pixel width and height to be a power of 2, we propose to divide the original summary in tiles whose dimensions correspond to the key-frames dimensions. This allows an accurate highlighting each key-frame while using precincts would have lead to an overlap of the neighbor key-frames. According to [13], the drawback of numerous small tiles is the block artifact it creates when decoding the image at low bitrate. However, in the case of a key-frame based image, such an artifact is obviously not a problem.

The meanings of each scene are annotated using a set of keywords from a predefined hierarchical lexicon. The annotations are saved in an MPEG-7 compliant XML file. As for the key-frame selection strategy, the quality of the scene annotation depends on the analysis system performances. In some cases, manual refinement or completion can be required.

## 2.2. System architecture

Figure 2 presents the proposed client-server system architecture that is based on the (EU-FP5) PRIAM project [23].

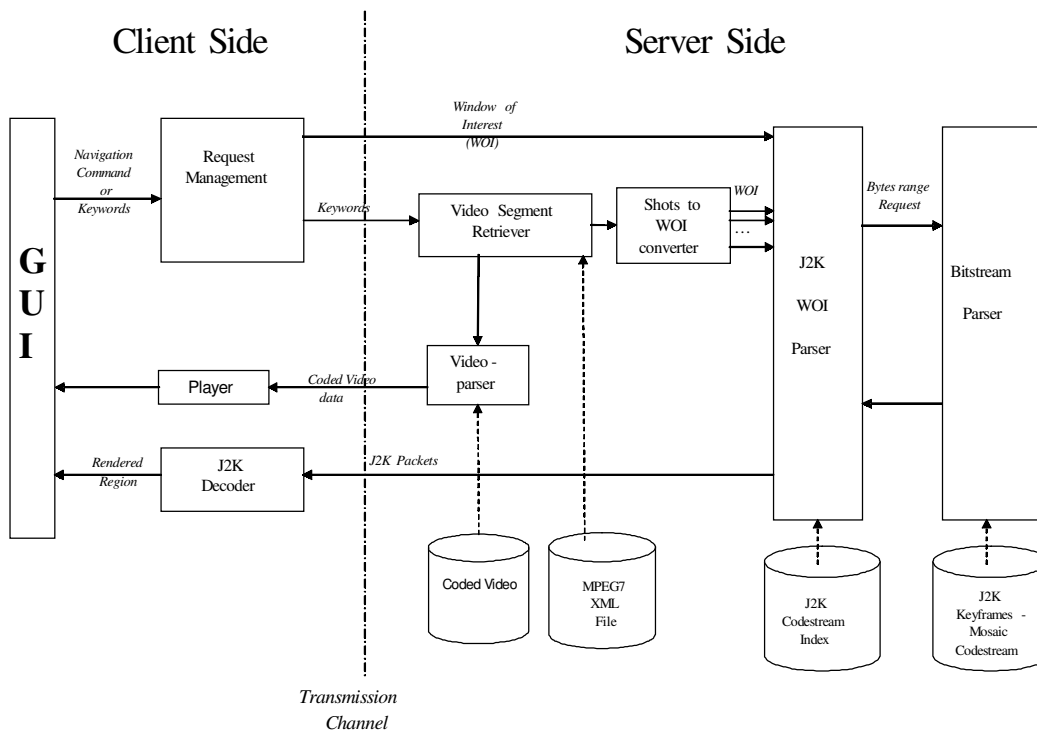


Figure 2. Client-server system architecture

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

- The navigation requests (zooming, panning etc) are translated at the client side into requests for Windows of Interest (WOI) [16]. Basically, a WOI specifies a spatial region, a quality level, a resolution level and the requested components:

$$WOI = [ x_0, y_0, x_1, y_1, \text{quality layer}, \text{resolution level}, \text{number of components} ],$$

With:

$$(x_0, y_0) = \text{Upper left corner coordinates of the requested spatial region,}$$

$$(x_1, y_1) = \text{Lower right corner coordinates of the requested spatial region.}$$

This first request translation is achieved by the request management module, which keeps information about the displayed data. At the server side, the JPEG 2000 WOI parser converts this WOI requests into the selection of relevant JPEG 2000 packets using the codestream index file. As mentioned in section 1, these packets contain additional data improving the quality of the requested regions once transmitted and decoded at the user's side. The packet selection is done taking the user's session history into account so as to avoid redundancies in the transmitted data.

- The retrieval queries, based on the keywords from the predefined lexicon, are linked at the server side with scenes indices, by the video segment retriever, which searches through the MPEG-7 annotation file of the video summary. The selected scenes are then associated to WOI's ('shots to WOI conversion'), as shown on Figure 3. The WOI's spatial region is defined by the scene's key-frame size and position in the summary. The corresponding WOI's specify the highest available quality and resolution levels so as to highlight key-frame of interest as much as possible compared to the non-relevant key-frames.



Table 1. Annotation lexicon used for the experiment

Event		Location	Object
Walking	Left	Left side	Man
	Right	Right side	Women
Browsing Adds		Shop Door	Group of people
Shop Enter		Inside shop	
Shop Exit			

From the constructed sequence, we extract key-frames describing each different scene and create the video summary following the method discussed in section 2.1.

Table 2. Video summary description

<b>Video frames dimensions</b>	640 x 480 [pixels <sup>2</sup> ]
<b>Number of key-frames</b>	6 x 5
<b>Video Summary size</b>	27 [MB]

Table 3. JPEG 2000 compressed summary description

<b>Number of resolution levels</b>	3
<b>Quality layers compression ratios</b>	100, 30, 1
<b>Tiles dimensions</b>	640 x 480 [pixels <sup>2</sup> ]
<b>Additional Marker</b>	SOP
<b>Compressed summary size</b>	5.9 [MB]

Table 2 gives the description of the original video summary and Table 3 lists the main compression options chosen for these preliminary tests. Figure 4 to Figure 7 show snapshots of the system in actions. Enhancing the visual quality and resolution of key-frames within the initial low quality overview clearly shows the temporal semantic links among the contents of these scenes.



Figure 4. First sequence overview: Low quality, Low resolution



Figure 5. "Browsing adds"



Figure 6. ' Group of people'



Figure 7. Zoom-in after requesting for ' Group of People'

## 4. CONCLUSIONS

A new method for building a scalable representation of key-frame-based storyboard is proposed by exploiting the powerful compression and scalability of JPEG 2000. Moreover, we have extended a JPEG 2000 Part 9 (JPEG 2000 Interactive Protocol) compliant platform to browse interactively the video summary, which could 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. This particularly suits surveillance applications where the context of detected events can be of high importance. The approach effectively exploits the user's interpretation capability while keeping the video summarization and description very simple.

The proposed system will be further tested on other long surveillance sequences and evaluated by a set of end-users in order to compare it with other video browsing systems. Moreover, new features like relevance feedback will be introduced in the system so as to adapt the system reaction to each end-user's behavior.

## ACKNOWLEDGEMENT

This work was partially supported by the EU FP5 Network of Excellence project SCHEMA ([www.schema-ist.org/SCHEMA/](http://www.schema-ist.org/SCHEMA/)) under project number IST-2001-32795. The video sequences used for the experiments described in section 3 are provided by the EU-FP5 project CAVIAR (IST-2201-37540)[24].

## REFERENCES

- [1] C. Kim and J.-N. Hwang, "Object-Based Video Abstraction for Video Surveillance Systems," *IEEE Trans. on Circuits and Systems for Video Technology*, **12** (12), pp. 1128-1138, Dec. 2002
- [2] H. Lee, A. Smeaton et al, "Implementation and Analysis of Several Key-frames-Based Browsing Interfaces to Digital Video," *Proc. of the 4<sup>th</sup> 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, "JPEG 2000: Standard for interactive imaging," *Proceedings of the IEEE*, **90**(8), pp. 1336-1357, Aug. 2002.
- [13] D. Taubman, "Remote Browsing of JPEG2000 Images," *Proc. of IEEE International Conference on Image Processing (ICIP)*, **1**, pp. 229-232, Sept. 2002.
- [14] D. Taubman, "High Performance Scalable Image Compression with EBCOT," *IEEE Trans. on Image Processing*, **9**, pp. 1158-1170, July 2000.

- [15] S. Deshpande and W. Zeng, "Scalable Streaming of JPEG2000 Images using Hypertext Transfer Protocol," *Proc. of ACM Multimedia*, pp. 281-372, Oct. 2001.
- [16] J. Meessen, T. Suenaga, M. Iregui Guerrero, C. De Vleeschouwer, and B. Macq, "Layered architecture for navigation in JPEG 2000 Mega-Images," *Proc. of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pp. 92-95, London, UK, April 2003.
- [17] 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>
- [18] D. Taubman and R. Prandolini, "Architecture, philosophy and performance of JPIP: internet protocol standard for JPEG 2000," *Proceedings of the International Symposium on Visual Communications and Image Processing (VCIP'2003)*, 2003.
- [19] 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.
- [20] 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.
- [21] JPSEARCH editors, "JPSEARCH scope and requirements," JPSEARCH project ISO/IEC 24800, available at URL: <http://www.jpeg.org>.
- [22] IBM Research, "VideoAnnEx Annotation Tool": available online at URL: [www.research.ibm.com/VideoAnnEx/](http://www.research.ibm.com/VideoAnnEx/).
- [23] EU IST FP5 project PRIAM (IST28646) "Platform for Real-Time and Interactive Access to Mega-images," <http://www.tele.ucl.ac.be/PROJECTS/PRIAM>.
- [24] EU IST FP5 project CAVIAR (IST-2001-37540) "Context Aware Vision using Image-based Active Recognition," <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>.