Content browsing and semantic context viewing through JPEG 2000-based scalable video summary

J. Meessen, L.-Q. Xu and B. Macq

Abstract: The paper presents a novel method and software platform for remote and interactive browsing of a summary of long video sequences as well as revealing the semantic links between shots and scenes in their temporal context. The solution is based on interactive navigation in a scalable mega image resulting from a JPEG 2000 coded key-frame-based video summary. Each key-frame could represent an automatically detected shot, event or scene, which is then properly annotated using some semi-automatic tools or learning methods. The presented system is compliant with the new JPEG 2000 Part 9 'JPIP – JPEG 2000 interactivity, API and protocols,' which lends itself to working under varying transmission channel conditions such as GPRS or 3G wireless networks. While keeping the advantages of a single 2D video summary, like the limited storage cost, the flexibility offered by JPEG 2000 allows the application to highlight interactively key-frames corresponding to the desired content first within a low-quality and low-resolution version of the full video summary. It then offers fine grain scalability for a user to navigate and zoom into particular scenes or events represented by the key-frames. This possibility of visualising key-frames of interest and playing back the corresponding video shots within the context of the whole sequence (e.g. an episode of a media file) enables the user to understand the temporal relations between semantically related events/actions/physical settings, providing a new way to present and search for contents in video sequences.

1 Introduction

With the advent of digital revolution, the availability of devices with more and more computing power and storage capacity, the rapid increase in network connectivity and bandwidth, alongside 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 visualise desirable contents [1]. A video content summary often takes the form of a 2D presentation layout on a visualisation interface; the summary is usually made up of selected frames, or key-frames, representing visually similar and/or semantically related data chunks, that is, shots or meaningful events. Depending on the media genre (e.g. news, sports, feature movies etc.) and applications, many different layouts for key-frames presentation are possible, as discussed in Lee *et al.* [2].

However, summarising the content of a long video sequence in this way for entertainment genres such as a feature movie or drama or for an unstructured surveillance video, the number of selected key-frames is still way too many. Two problems are ensued. First, the semantic story structures will be largely buried in the numerous images displayed. Secondly, in the case of a user accessing/browsing the video summary stored on a remote server, the transmission over the network of a large number of key-frames is a potential bottleneck. Today, there exist a number of different approaches to addressing these issues. Building condensed and semantically relevant video summaries have been seen in the work by Yeung and Yeo [3] and by Chiu *et al.* [4]. However, although these summaries present a good overview of the content in a sequence, they tend to provide a user with only one predefined semantic level illustration. If it does not happen to correspond to the user required semantic level, this solution is not efficient. Consequently, despite that this type of summary allows limited storage, with only selected key-frames being required available, it suffers from a lack of flexibility.

Hierarchical clustering and presentation is the common approach to browse either one video file [5–7] or a database of video sequences [8]. In the case of browsing one video sequence, the shots are clustered with other shots sharing certain similar low-level visual descriptors/features. As illustrated in Fig. 1, these clusters are then further grouped to form larger clusters based on higher and more abstract content similarities, and the process continues up to potentially very high semantic descriptions. Hierarchical browsing of a database of video sequences relies on the same principle. The sequences are grouped in with each other in different levels, using semantic similarity measures from low level (visual features like colours, texture, shapes etc.) to very high level (subjective concepts). In both cases, the browsing system starts with presenting the highest semantic level of the database, or

IEE Proceedings online no. 20050066

doi:10.1049/ip-vis:20050066

Paper first received 3rd March and in revised form 21st June 2005

J. Meessen is with Multitel asbl, Avenue Copernic 1, Mons B-7000, Belgium

L.-Q. Xu is with BT Research and Venturing, Adastral Park, Ipswich IP5 3RE, UK

B. Macq is with the Communication and Remote Sensing Laboratory, Université catholique de Louvain, Place du Levant 2, Louvain-la-NeuveBelgium

E-mail: jerome.meessen@multitel.be

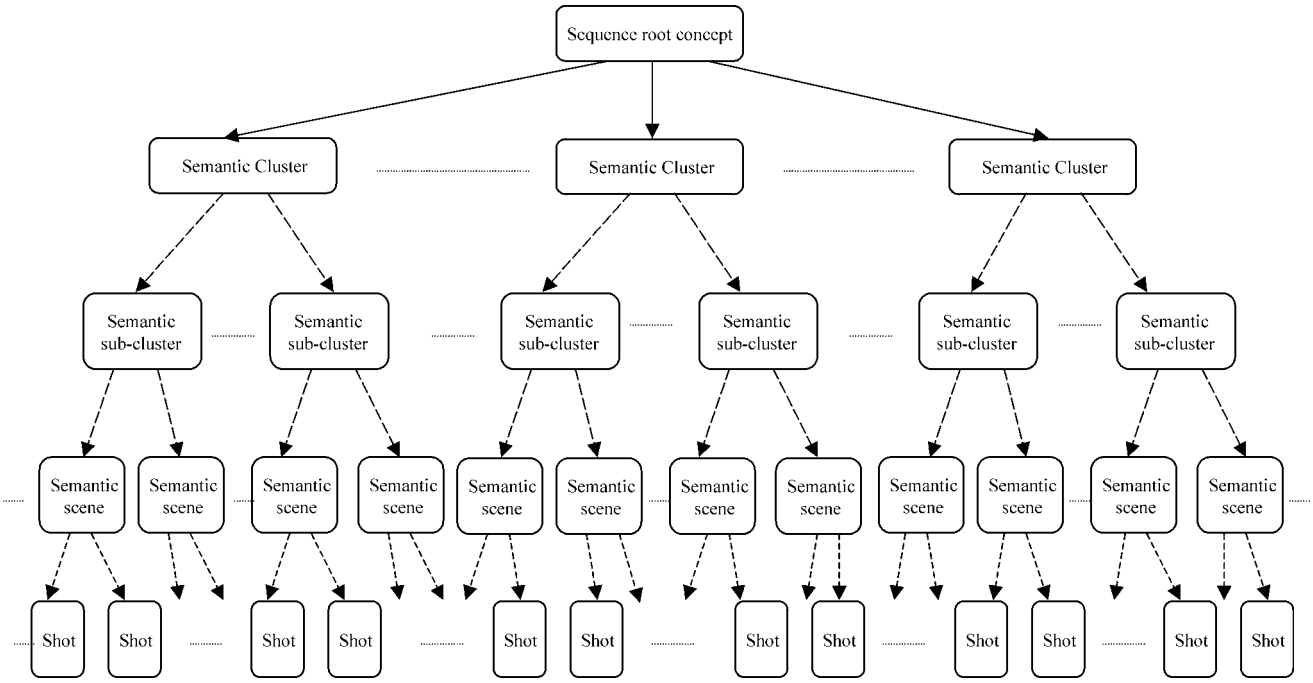

Fig. 1 *Illustration of the hierarchical video semantic structure*

sequence, it then allows a user to delve into the lower clusters along the tree branches in order to retrieve a certain sequence or scene of interest. In particular, the shot clustering and browsing methods [9, 10] are especially interesting, as they are evaluated with respect to the amount of transmitted data required at each user retrieval request, which is often a critical requirement. Interactive retrieval operations like hierarchical browsing contribute to a faster and deeper understanding of video content. Consequently, such operations must be considered when building a 2D video summary.

These are some of the efficient solutions to acquire a quick overview of a video sequence and to find a particular scene of interest. However, they do not have provisions for contextual visualisation of the links between semantically related scenes, that is, 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?' and so on.

In this paper, we focus on investigating a flexible video content visualisation mechanism that helps a user to understand the underlying semantic structure of a video sequence, that is, to establish relations between semantically similar scenes and highlight them within the context of the whole sequence when browsing only through one 2D summary of the video sequence. Rather than propose a complex shot clustering strategy or storyboard layout, we aim to exploit the user's intelligence by providing him/her with a set of tools enabling interactive browsing and navigation in a remotely stored scalable summary. In the quest for solutions to bridging the semantic gap between automatically computed low-level features and high-level concepts, we choose to resort to tap end users' unique interpretation skills.

2 System framework

This section starts with an overview of the rationale of the proposed method with a view to interactive video content browsing and semantic context viewing, it then proceeds to discuss the two core components. The first one is about the creation of a scalable video content summary with suitable semantic annotation of its content units, taking advantage of several unique features furnished by JPEG 2000 standard. The second is on the extension of the client–server system platform previously developed for mega-image navigation to host and transmit the generated video summary, facilitating semantic-based search and browsing of the video content by end users.

2.1 Overview of the method

The core idea of the proposed solution rests upon exploiting the powerful features of data compression and scalable representation furnished by JPEG 2000, the new standard for still image compression [11], in order to produce scalable key-frame-based summaries of a video sequence, while at the same time allowing for semantics-based video clip queries. This scalable representation enables interactive browsing of a low storage cost key-frame-based video summary.

JPEG 2000 offers a highly scalable representation of the compressed image, in terms of image components, spatial access, resolution and quality [12]. This is particularly suited to browsing, or navigating within, very large images (the so-called 'mega images') as discussed in Meessen *et al.* [13] and Taubman [14]. Colour components are coded separately, so that they can be accessed independently. The resolution scalability is due to a discrete wavelet transform (DWT). The quality scalability is achieved by the optimal code-stream decomposition in different quality layers that are traditionally created through rate-PSNR optimisation [15]. The spatial access to different regions of the image can be obtained, thanks to the several different mechanisms available. The source image can be tiled in smaller images, or tiles, which are compressed independently and the contributions of which are signalled in the 'code-stream', the coded data stream. Another way to access spatial regions is provided by the precinct partitioning of the DWT coefficients [16]. Within the code-stream, a packet contains the data corresponding to one precinct at a given resolution level and from one quality layer. The code-stream is further provided with marker segments signalling

all these packets so that each one can be retrieved at random.

In this paper, 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 (alongside a corresponding but separately accessed semantic annotation data file), this standardised communication between a server and client provides the tools for exploiting the JPEG 2000 scalability, and accessing interactively many different versions of the storyboard over a network. This enables the system to offer interactive semantic browsing using only one representation of the video content. JPIP uses a 'code-stream index,' which basically details the coding options used for creating the code-stream, or the number of quality layers, resolution levels and so on, and references the position of the JPEG 2000 code-stream elements like headers, tiles, precincts, packets and so forth. More details about the structure of the code-stream 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.

The annotation of video content for the current studies of video summary is based on MPEG-7 multimedia description schemes [19]. The input video sequence is first partitioned into temporally smaller segments such as shots or events (a group of adjacent shots) based on comparing similarities in low-level visual features [20]. After this temporal decomposition of a video sequence into segments, a semi-automatic inference method or an interactive annotation tool can be used to annotate the content of each segment taking account of both the dynamic characteristics (actions) of the temporal segment and the visual appearance (background scene, objects etc.) of the representative key-frames. In the end, an MPEG-7 compliant XML description file is obtained, specifying, 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 [19]. The annotation of video segments in this manner allows translating content-based queries into image-oriented requests matching. It is worth pointing out that the JPEG standardisation group has recently launched a new work item, called JPSEARCH, intending to extend the JPEG 2000 XML metadata fields in order to enable new content-based applications like image retrieval [21]. As our annotations structure is based on XML schemes, it could be easily modified to be in compliance with this emerging standard.

2.2 Scalable key-frame-based video summary with annotation

Fig. 2 depicts schematically the workflow used to create the coded key-frame-based video summary with desired annotation. The original MPEG compressed video sequence is first segmented into shots, and for each shot one key-frame is selected to represent its visual content. The representation scheme can be extended to a group of shots, or subscene or scene, by merging successively adjacent key-frames bearing certain degree of visual similarity in order to avoid the redundancy in displayed visual content, as in the case of an excerpt from the film 'A beautiful mind' to be discussed in the experimental section. This can be done easily through a video editing tool, or better using automated images matching techniques according to certain low-level visual features. The final set of key-frame images is then concatenated in raster scanning order to compose a large mosaic image, referred to as 'video-image' thereafter, which is then JPEG 2000 compressed to output two binary data files: the JPEG 2000 code-stream and its associated 'code-stream index' file as defined in JPEG 2000 Part 9 'JPIP'. The compression of

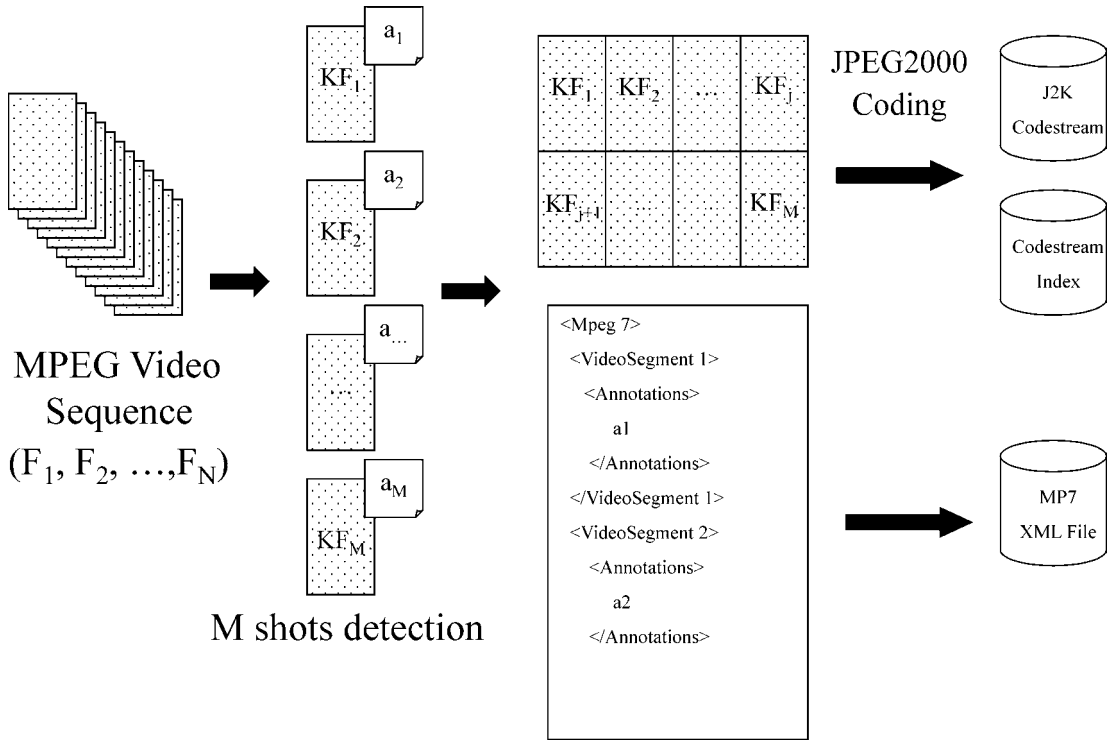

Fig. 2 *Creation of a scalable key-frame-based annotated video summary, which is coded by JPEG 2000 in preparation for interactive system access*

the video-image is performed with at least two quality layers and several resolution levels, so that the target image regions can be highlighted within the whole video-image, or the regions corresponding to the key-frames of interest are decoded and rendered with better visual quality when compared with the rest of the video-image. Moreover, the dimensions of the JPEG 2000 tiles and precincts are chosen such that each key-frame can be accessed separately. As the JPEG 2000 coding standard requires that the width and height of a precinct are a power of 2, we rather propose to divide the original video-image into tiles whose dimensions correspond to the key-frames' dimensions. This allows an accurate highlighting of each key-frame in the display, whereas using precincts would have led to an overlap of neighbouring key-frames as shown in Figs. 3 and 4. According to Taubman [14], the main drawback of numerous small tiles is the blocking artefact thus introduced when decoding the image at low bitrate. However, in the case of our video-image that is itself made up of blocks of key-frame images, such an artefact is obviously not a problem.

Moreover, as tiles are coded independently, the code-stream is basically structured as a main header followed by a sequence of tile-streams in raster scan order, as illustrated in Fig. 5. The tiles are then easily accessible within

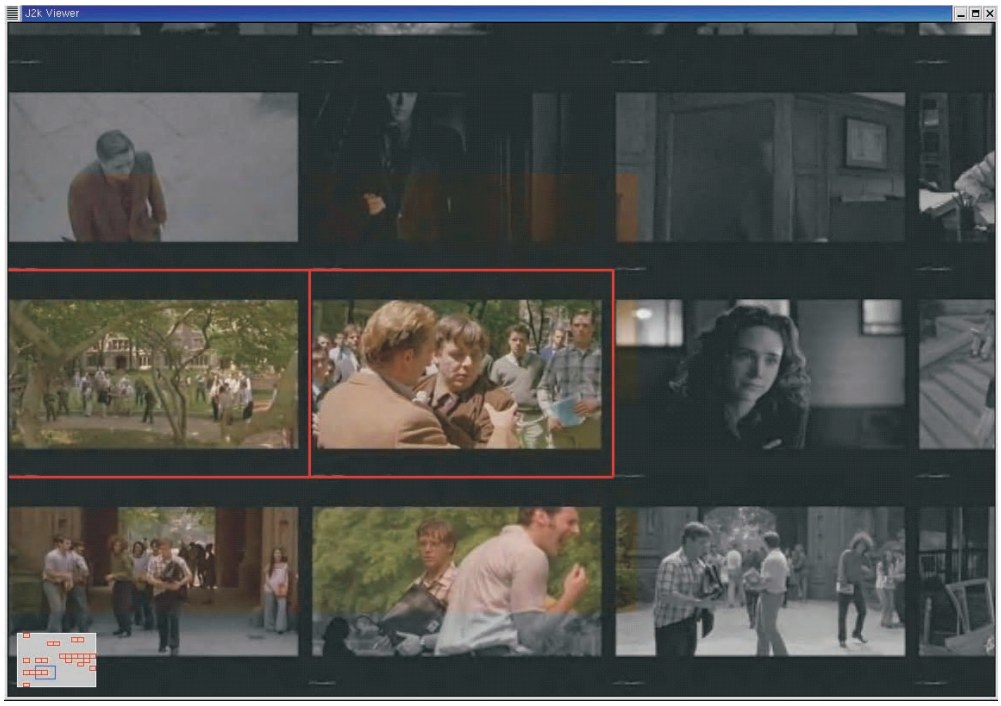

Fig. 3 *Highlighting of two consecutive keyframes using precinct-based spatial access*
Non-required overlap of the lower key-frames is observed

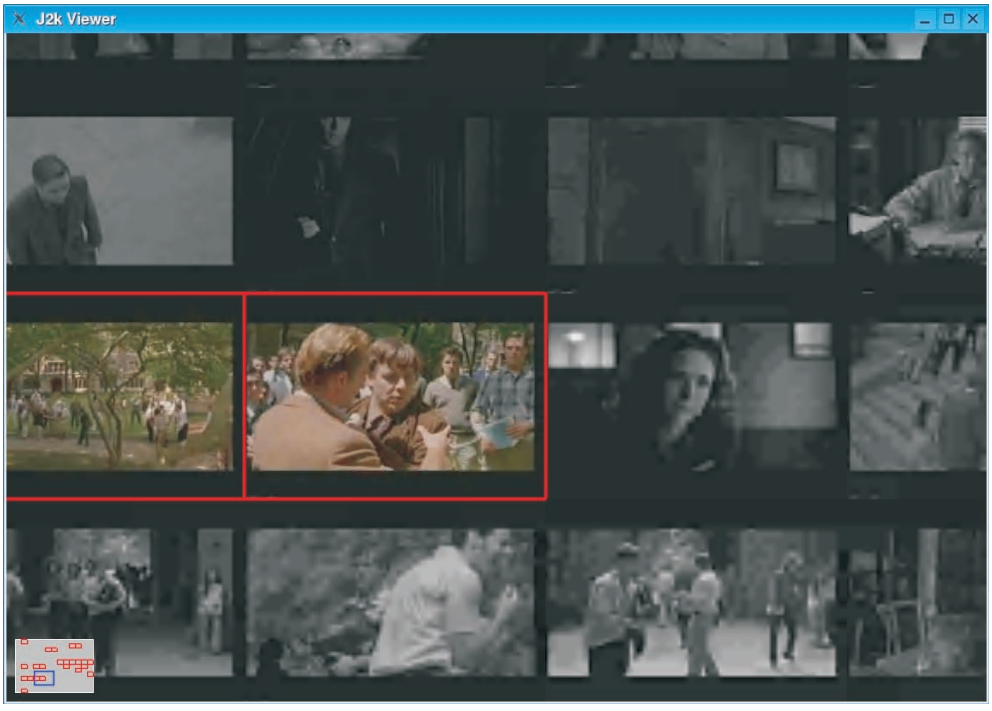

Fig. 4 *Highlighting two consecutive key-frames using tile-based spatial access*

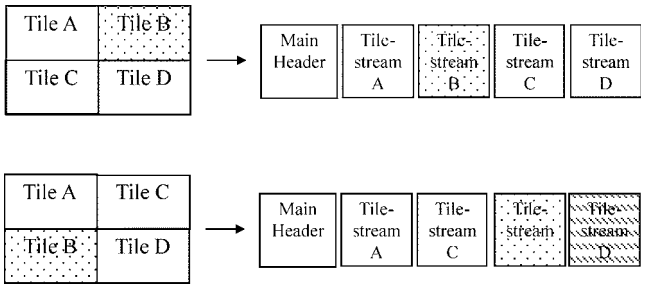

Fig. 5 *Raster scan order of the tile-streams in a JPEG 2000 codestream*

Dynamic arrangement of the tiles straightforward

the code-stream. As we are focusing on showing temporal relations between semantically similar key-frames, the time-based mosaic ordering has been preferred in the current implementation. However, this direct access to the tiles allows dynamic key-frame ordering based on other criteria, for example, on their relevance to the user request. This would require only minor signalling modifications to each displaced tile-stream.

Having described the way how the summary video-image is created and encoded, we now turn to the issue how to annotate the video shots/scenes as represented by the key frames in the representation such that the access to the desired content can be fulfilled using text-based semantic search. Automatic annotation of images through machine learning and language translation techniques is still an opening research issue, and it works for large image collections with clear categorical difference [22, 23]. For the video programme of limited number of key-frames, we choose to use semi-automatic annotation tools to manually annotate each shot/scene. One choice is the Professional Annotation Client described in Section 4.3 of Xu *et al.* [24] and developed in the recently completed EU FP5 project BUSMAN [25]. In this case, a simple and intuitive keyword-based approach based on the 'StructuredAnnotation' MPEG-7 descriptor is adopted. This descriptor allows storing textual annotations in terms of seven basic semantic concepts, 'animate objects – people and animals' (Who), 'objects' (WhatObject), 'actions' (WhatAction), 'places' (Where), 'time' (When), 'purposes' (Why) and 'manner' (How). An annotator can associate any number of these semantic descriptors to a given shot. Another option is the use of 'VideoAnnEx' annotation tool from IBM Research [26]. The meanings of each shot are annotated using a set of keywords from a predefined hierarchical lexicon, reflecting, respectively, the static scene descriptions, key object descriptions, event descriptions and so on. In both the above annotation tools, the shot segmentation results can be manually edited by splitting and merging shots to overcome the segmentation errors. And the annotations are saved in an MPEG-7 compliant XML file.

Let us note that a semantic content description, that is, using the MPEG-7 'Semantic' descriptor, rather than the earlier mentioned structural description is totally compatible with the proposed system.

2.3 System architecture

In the previous section, we have presented our proposed method for creating a scalable video content summary made of annotated key-frames. In this section, we present the proposed navigation and retrieval system. Fig. 6 presents the client–server system architecture, which extends the work performed in the IST PRIAM project [27]. We consider three types of client requests: the navigation requests for browsing around the video, the retrieval requests searching for desired semantic events/scenes and video playing requests to view the dynamic video content of selected shots.

2.3.1 Navigation requests: The navigation requests (zooming, panning etc.) are translated at the client side into requests for windows of interest (WOI) [13]. Basically, a WOI is a vector specifying a spatial region, a quality level, a resolution level and the requested

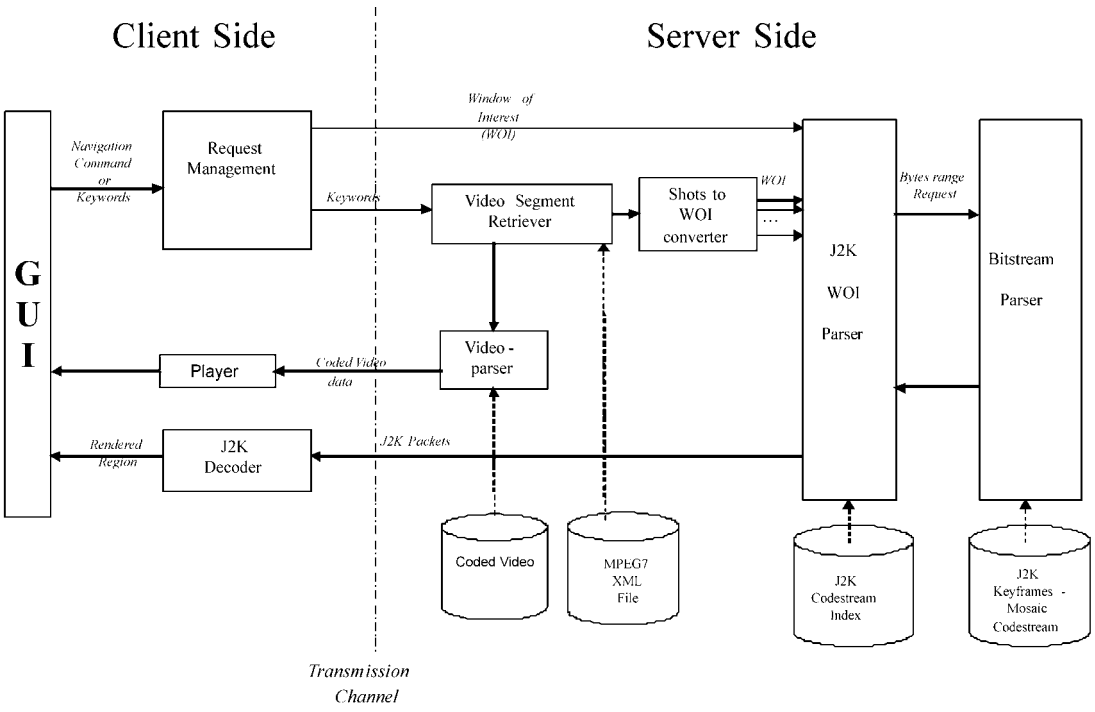

Fig. 6 *Proposed client–server system architecture for effective presentation of a video summary towards interactively browsing, search and playing desired video events*

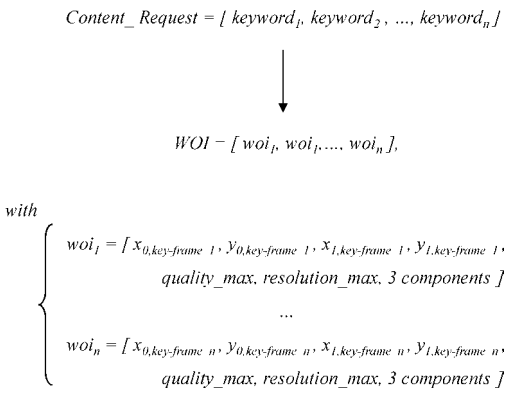

Fig. 7 *Translation of content-based requests in terms of keywords into a request for WOI in the JPEG 2000 compressed 'code-stream'*

components

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

with

(x_0, y_0) = Top left corner coordinates of the requested spatial region

(x_1, y_1) = Bottom 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 request into the selection of relevant JPEG 2000 packets using the code-stream index file. As mentioned in Section 2.1, 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. As discussed in Meessen *et al.* [13], such a layered platform allows different client–server configurations in line with their respective processing and memory resources.

2.3.2 Retrieval requests: The retrieval queries are conducted using the keywords from a predefined lexicon, which are linked at the server side with scenes indices of semantic annotation, by the video segment retriever. The retrieving module searches through the MPEG-7 annotation file of the video summary for the scenes corresponding to the query keywords. The selected scenes are then associated with WOI's ('shots to WOI conversion'), as shown in Fig. 7. The WOI's spatial region is defined by the scene's key-frame size and position in the summary. The corresponding WOI's specify all colour components at the highest quality and resolution levels available, so as to highlight the key-frames of interest as much as possible compared with the non-relevant key-frames.

2.3.3 Video playing requests: The video playing requests are necessary if the user after browsing through the highlighted static images (key-frames of the shots/scenes) would like to view the video clips (dynamic contents) that they depict. Thanks to the MPEG-7 complied content description file, the system retrieves the temporal boundaries of the required shot, which are used by a video player module implemented to allow the user to play this particular scene of interest.

3 Experimental studies

In this section, we discuss the application scenarios of the proposed system and present the experimental results obtained with respect to three test video sequences.

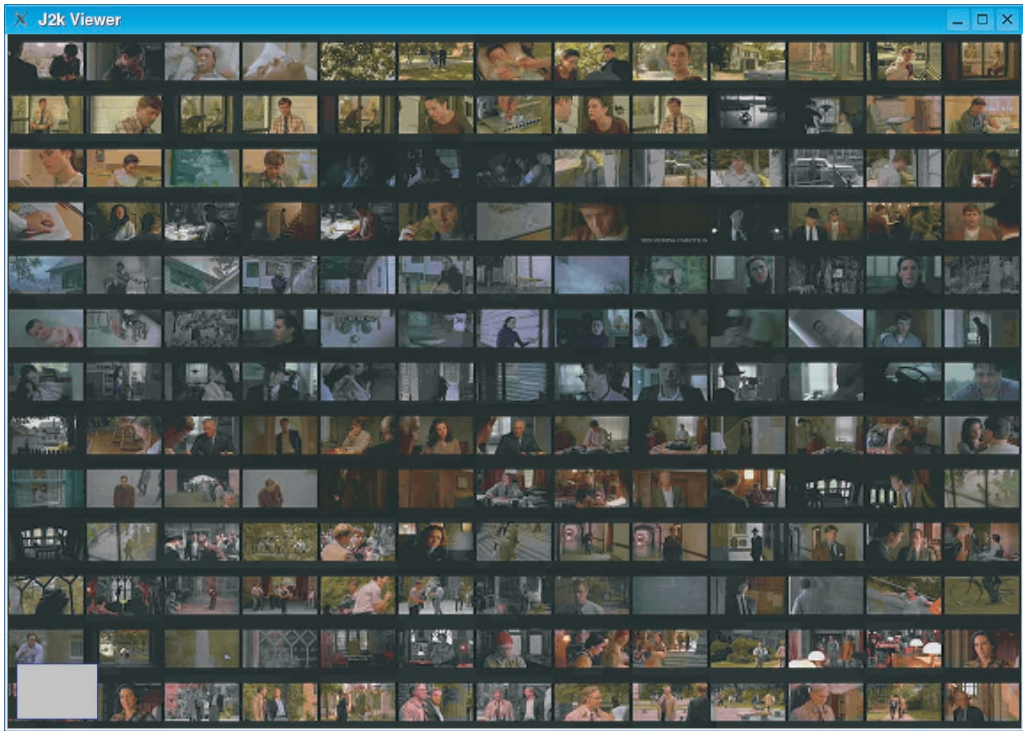

Fig. 8 *Coloured and low-quality background of the excerpt from 'A beautiful mind'*

Fig. 9 *Highlighting of key-frames corresponding to 'crying' scenes, within a coloured low-quality background*
The distinction between highlighted and non-highlighted key-frames would not be possible without additional red rectangles

3.1 Typical scenario

A typical application scenario is described as follows. To start with, the GUI of the system's client displays an overview of the JPEG 2000 coded entire video-image, or static video summary, which is obtained by decoding only the lowest resolution and quality levels.

There are two different ways to present the first low-quality overview of the summary. On the one hand, only a greyscale version can be used. Although only limited information is available for this first overview of the key-frames, this ensures high contrast between highlighted and non-highlighted key-frames when answering to a client request. On the other hand, a more compressed coloured version of the summary provides more visual information at this first overview. However, after retrieving scenes of interest, the highlighted key-frames are not easily distinguishable from the non-relevant key-frames.

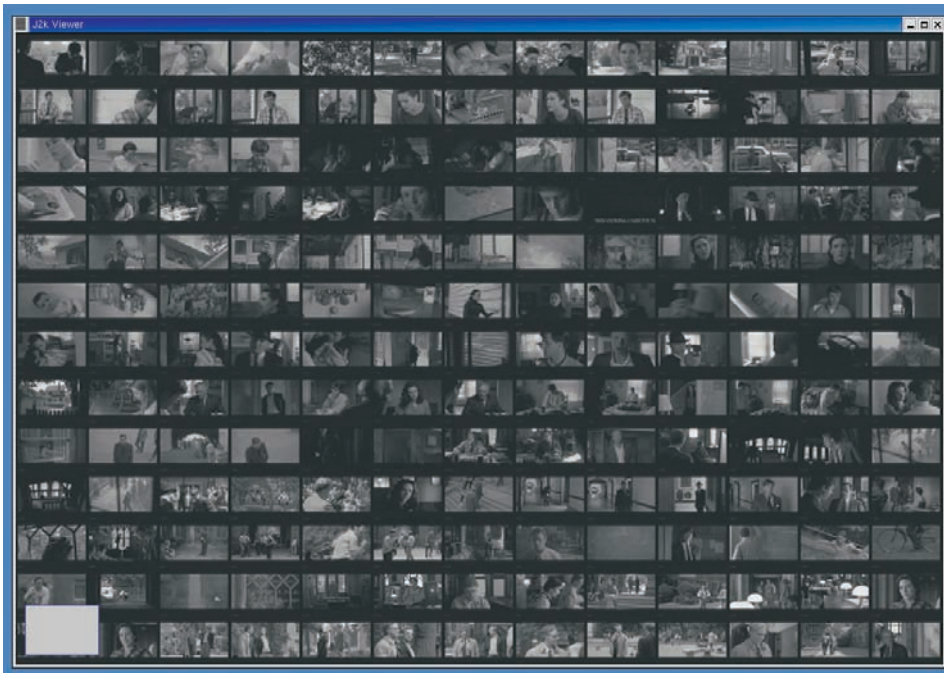

Fig. 10 *At the start of an interactive user browsing and search session, the interface shows a greyscale low-quality video summary of the video excerpt from the movie 'A beautiful mind'*

Table 1: Characteristics of the three test video sequences and the associated key-frame-based video summary

Test videos	Notting Hill	A beautiful mind	Surveillance video
Video length, min	20	40	~10
Key-frame size	352 × 288	352 × 240	640 × 480
Number of key-frames	240 = 16 × 15	169 = 13 × 13	30 = 6 × 5
Summary dimensions	5632 × 4320	4576 × 3120	3840 × 2400
JPEG 2000 compressed summary size, KB	1400	837	902

Additional graphical tricks are necessary to ensure a clear highlighting of the retrieved key-frames, as shown in Figs. 8 and 9. In our case, as the focus is on highlighting related scenes and showing their temporal distribution, rather than providing as much information as possible for the non-relevant key-frames, the greyscale version of the summary is preferred (Fig. 10).

The user can then select the keywords from the annotation dictionary and request the server to search for and present certain desired and more detailed contents of the video. The relevant key-frames as retrieved by the server are subsequently highlighted by the GUI, or supplied with more bits of colour information. It should be emphasised that enhancing the visual quality and resolution of key-frames within the initial low-quality greyscale overview clearly shows the temporal semantic links among the contents of these selected 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.

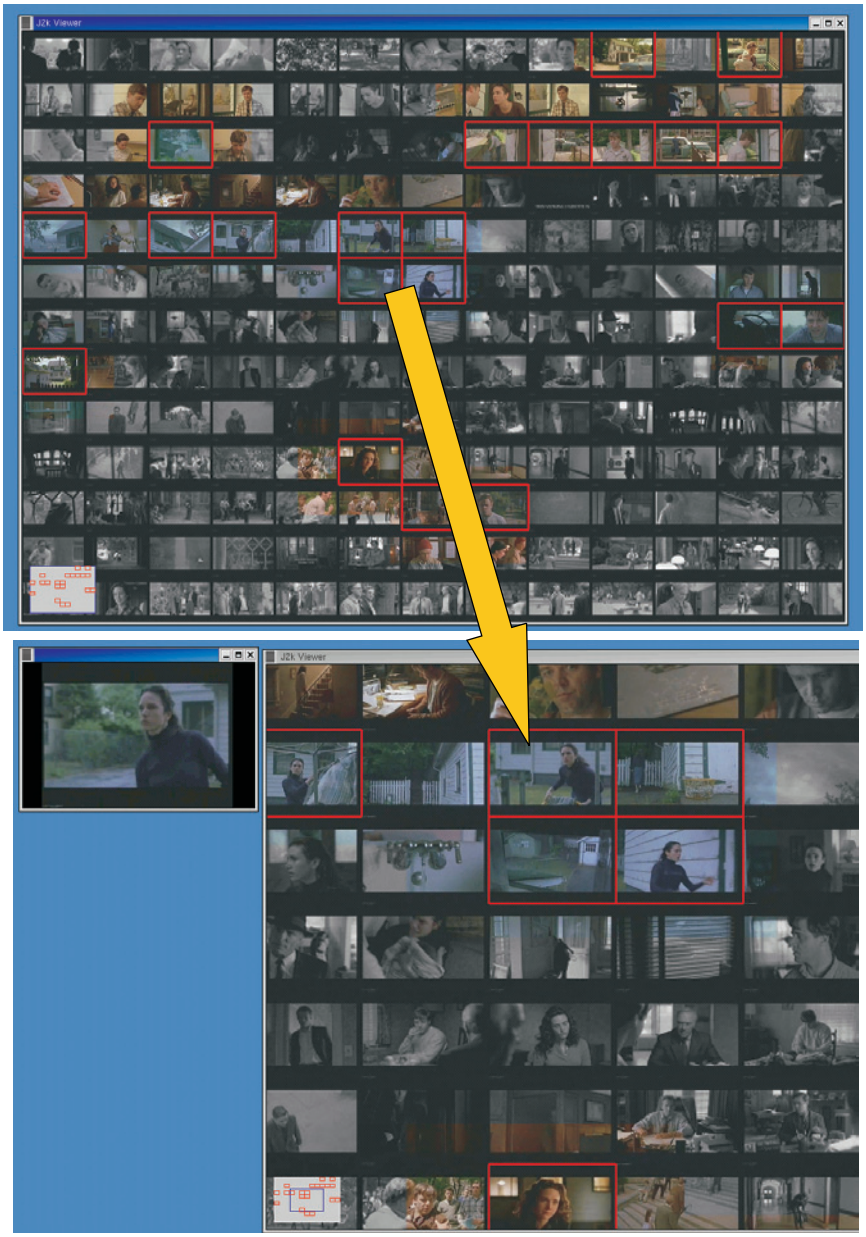

Fig. 11 *Screen shots of the proposed system on the video excerpt from the movie 'A beautiful mind'*
Key-frames corresponding to the retrieval query 'Urban' are highlighted within the lower quality overview (top)
JPEG 2000 allows the panning and zoom

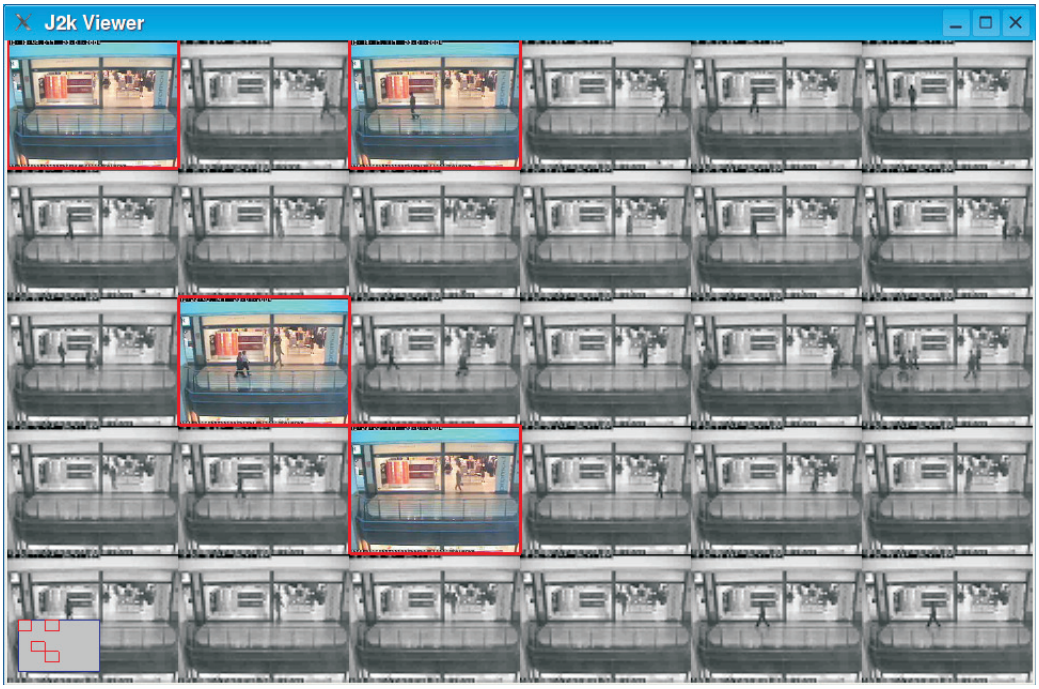

Fig. 12 *Snapshot of the system when browsing the surveillance video summary*

The keyframes corresponding to the request 'People walking on the left side' are highlighted with higher resolution, quality and all colour components

3.2 Experiments

To evaluate the performance and functionality of the prototype system, experiments are performed on three test video files, including one excerpt from the British comedy 'Notting Hill,' another excerpt from the feature movie 'A beautiful mind' and a third one from a publicly available benchmarking video surveillance sequence [28]. Table 1 specifies the details of the three test video sequences together with the attributes of their corresponding key-frame-based video-image and the respective JPEG 2000 compressed version. As the given surveillance sequence is very short (10 min), a high number of key-frames has been voluntarily selected by hand, corresponding to an average of three key-frames per minute. Although not all these key-frames correspond to actual events to be observed in the sequences, this experiment sets to demonstrate the usability of the system in this application field.

Snapshots of the system in action are shown in the following figures. Fig. 10 presents the greyscale and lowest quality overview on display at the start of a user interactive browsing and search session of the excerpt from the movie 'A beautiful mind'. This first version of the summary mainly helps the user to obtain a global overview of the video content and the succession of stories before formulating a specific semantic-based request. In Fig. 11 (top), selected key-frames (shots) are popped out after such a request asking for 'urban' scenes. These key-frames are coloured and provided with full resolution and quality so that they are highlighted compared with the rest of the overview image. The inset in the bottom-left indicates the current viewing window and the positions of highlighted shots. The scalability of JPEG 2000 allows low-cost panning and zooming to explore particular parts of the highlighted video summary, before asking for playing a shot of interest in the left panel, see Fig. 11 (bottom). Fig. 12 presents a snapshot of the system when browsing through the test video surveillance sequence, in which the key-frames corresponding to the request for the appearance of 'People walking on the left side' are highlighted with higher image resolution and higher quality.

4 Conclusions

A new method and software platform for building a scalable representation of key-frame-based video storyboard have been investigated, exploiting the powerful compression and scalability features of JPEG 2000. The proposed approach benefits from the advantages of a single light 2D video content summary, while providing interactive and intuitive browsing features. Specifically, we have extended a JPEG 2000 Part 9 'JPIP' compliant platform to browse interactively the key-frame-based summary of a long video sequence, which could be accessed under different networking conditions or processing resources. Using the MPEG-7 description scheme to annotate the semantic content of each shot/scene of the video sequences, the proposed system allows the user to search for desired events and scene and visualise the links between semantically similar scenes; it therefore provides a new way to understand long video sequences. The prototype system has been tested using two long movie excerpts and a surveillance video sequence.

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

Further work to extend the existing features is planned, including the provision of functionalities enabling queries by example search for desired content. Investigation into advanced semantic search mechanism like hidden annotation and relevance feedback will be carried out so as to adapt the system reaction to each end-user's behaviour, using the dynamic key-frame arrangement discussed in Section 2.2. The system will also be extended to the case where the compressed video sequences are locally stored and the mosaic can be built with references to their frames directly.

5 References

1 Smeaton, A.: 'Challenges for the content-based navigation of digital video in the Fishlar digital library'. Proc. Int. Conf. on Image and Video Retrieval (CIVR'02), London, UK, July 2002, pp. 215–224

2 Lee, H., Smeaton, A., Berrut, C., Murphy, N., Marlow, S., and O'Connor, N.: 'Implementation and analysis of several key-frames-based browsing interfaces to digital video'. Proc. 4th European Conf. on Digital Libraries (ECDL'00), Lisbon, Portugal, September 2000, pp. 206–218

3 Yeung, M., and Yeo, B.: 'Video visualization for compact presentation and fast browsing of pictorial content', *IEEE Trans. Circuits Syst. Video Technol.*, 1997, **7**, (5), pp. 771–785

4 Chiu, P., Girgensohn, A., and Liu, Q.: 'Stained-glass visualization for highly condensed video summaries'. Proc. IEEE Int. Conf. on Multimedia and Expo (ICME'04), Taipei, Taiwan, June 2004, vol. 3, pp. 2059–2062

5 Ferman, A.M., and Tekalp, A.M.: 'Two-stage hierarchical video summary extraction to match low-level user browsing preferences', *IEEE Trans. Multimedia*, 2003, **5**, (2), pp. 244–256

6 Shipman, F., Girgensohn, A., and Wilcox, L.: 'Generation of interactive multi-level video summaries'. Proc. ACM Conf. on Multimedia, Berkeley, USA, November 2003, pp. 392–401

7 Fan, J., Elmagarmid, A., Zhu, X., Aref, W., and Wu, L.: 'Classview: hierarchical video shot classification, indexing, and accessing', *IEEE Trans. Multimedia*, 2004, **6**, (1), pp. 70–86

8 Taskiran, C., Chen, J.-Y., Albiol, A., Torres, L., Bouman, C., and Delp, E.: 'Vibe: a compressed video database structured for active browsing and search', *IEEE Trans. Multimedia*, 2004, **6**, (1), pp. 103–118

9 Smith, J.R.: 'Videozoom spatio-temporal video browser', *IEEE Trans. Multimedia*, 1999, **1**, (2), pp. 157–171

10 Doulamis, A., and Doulamis, N.: 'Optimal content-based video decomposition for interactive video navigation', *IEEE Trans. Circuits Syst. Video Technol.*, 2004, **14**, (6), pp. 757–775

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

12 Taubman, D., and Marcellin, M.: 'JPEG 2000: standard for interactive imaging', *Proc. IEEE*, 2002, **90**, (8), pp. 1336–1357

13 Meessen, J., Suenaga, T., Guerrero, M.I., De Vleeschouwer, C., and Macq, B.: 'Layered architecture for navigation in JPEG 2000 mega-images'. Proc. 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'03), London, UK, April 2003, pp. 92–95

14 Taubman, D.: 'Remote browsing of JPEG 2000 images'. Proc. IEEE Int. Conf. on Image Processing (ICIP'02), Rochester, New York, September 2002, vol. 1, pp. 229–232

15 Taubman, D.: 'High performance scalable image compression with EBCOT', *IEEE Trans. Image Process.*, 2000, **9**, (7), pp. 1158–1170

16 Deshpande, S., and Zeng, W.: 'Scalable streaming of JPEG 2000 images using hypertext transfer protocol'. Proc. ACM Conf. on Multimedia, Ottawa, Ontario, Canada, October 2001, pp. 281–372

17 JPIP (Eds.: 'JPEG 2000 image coding system – Part 9: Interactivity tools, APIs and protocols – Final Committee Draft 2.0', December 2003), available at http://www.jpeg.org/public/fcd15444-9v2.doc

18 Taubman, D., and Prandolini, R.: 'Architecture, philosophy and performance of JPIP: internet protocol standard for JPEG 2000', *Proc. SPIE–Int. Soc. Opt. Eng.*, 2003, **5150**, pp. 649–663

19 Salembier, P., and Smith, J.: 'MPEG-7 multimedia description schemes', *IEEE Trans. Circuits Syst. Video Technol.*, 2001, **11**, (6), pp. 748–759

20 Sikora, T.: 'The MPEG-7 visual standard for content description – An overview', *IEEE Trans. Circuits Syst. Video Technol.*, 2001, **11**, (6), pp. 696–702

21 JPSEARCH (Eds.): 'JPSEARCH scope and requirements'. JPSEARCH project ISO/IEC 24800, available at http://www.jpeg.org

22 Duygulu, P., Kobus, B., de Freitas, J.F.G., and Forsyth, D.A.: 'Object recognition as machine translation: learning a lexicon for a fixed image vocabulary'. Proc. 7th European Conf. on Computer Vision (ECCV'020), 2002, vol. IV, pp. 97–112

23 Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., and Jordan, M.I.: 'Matching words and pictures', *J. Mach. Learn. Res.*, 2003, **3**, pp. 1107–1135

24 Xu, L-Q., Villegas, P., *et al.*: 'A user-centred system for end-to-end secure multimedia content delivery: from content annotation to consumer consumption'. Proc. Int. Conf. on Image and Video Retrieval (CIVR'04), Dublin, Ireland, July 2004

25 (EU IST BUSMAN Project: : 'Bringing user satisfaction to media access networks', 2001–2004, available at http://www.ist-busman.org

26 Smith, J.R., and Lugeon, B.: 'A visual annotation tool for multimedia content description'. Proc. SPIE Photonics East, Internet Multimedia Management Systems, November 2000, (see also, IBM Research 'VideoAnnEx Annotation Tool', available at http://www.research.ibm.com/VideoAnnEx)

27 EU IST FP5 project PRIAM (IST28646): 'Platform for real-time and interactive access to mega-images', available at http://www.tele.ucl.ac.be/PROJECTS/PRIAM

28 EU IST FP5 project CAVIAR (IST-2001-37540): 'Context aware vision using image-based active recognition', available at http://homepages.inf.ed.ac.uk/rbf/CAVIAR/