Noname manuscript No.
(will be inserted by the editor)

Embedded Indexing in Scalable Video Coding

Nicola Adami · Alberto Boschetti · Riccardo Leonardi · Pierangelo Migliorati

Received: date / Accepted: date

Abstract Effective encoding and indexing of audiovisual documents are two key aspects for enhancing the multimedia user experience. In this paper we propose the embedding of low-level content descriptors into a scalable video-coding bitstream by jointly optimizing encoding and indexing performance. This approach provides a new type of bitstream where part of the information is used for both content encoding and content description, allowing the so called "Midstream Content Access". To support this concept, a novel technique based on the appropriate combination of Vector Quantization and Scalable Video Coding has been developed and evaluated. More specifically, the key-pictures of each video Group Of Pictures (GOP) are encoded at a first draft level by using a suitable visual-codebook, while the residual errors are encoded using a conventional approach. The same visual-codebook is also used to encode all the key-pictures of a video shot, where boundaries are dynamically

All the authors:

DEA-SCL, University of Brescia, Via Branze 38, 25123, Brescia, Italy

Tel.: +39 030 371 5902

Fax: +39 030 380014

E-mail: {firstname.lastname}@ing.unibs.it

estimated. In this way, the visual-codebook is freely available as an efficient visual descriptor of the considered video shot. Moreover, since a new visual-codebook is introduced every time a new shot is detected, also an implicit temporal segmentation is provided.

1 Introduction

The efficient encoding and indexing of audiovisual sequences are both important aspects for improving the usability and capability of modern multimedia systems. Traditionally, content coding and content indexing have been mainly studied as separate problems. At our best knowledge, the only relevant joint approach is described in the part of the MPEG-4 standard dealing with media object. Anyway its application is limited to synthetic content due to the difficulties of automatically and precisely identify foreground/background audiovisual objects and their relationship. Traditional image and video coding approaches address the problem of Rate-Distortion (RD) optimization, trying to obtain the maximum quality (in terms of SNR) at a given bit-rate. An evolution of these coding schemes tries to obtain the scalability of the coded bitstream without affecting the coding performance, in terms of RD [21], [23], [2]. Scalable Image Coding (SC) methods, such as for example JPEG2000 (JP2K) [22], generate a bitstream with the possibility of extracting decodable sub-streams corresponding to a scaled version, i.e., with a lower spatial resolution and/or a lower quality, of the original image. Moreover, this is achieved providing coding performance comparable with those of single point coding methods and requiring a very low sub-stream extraction complexity, actually comparable with read and write operations.

Beside, the main objective of content based analysis is to extract indexes, at different semantic level, which can be used for filling the semantic gap between the user expectance and the actual results provided by fast content browsing and content retrieval applications.

In this context, particular attention is given to the extraction of low-level descriptors of the considered audio-visual information [10], [6], [11], which, for example, can be used in query by example operation or as a base for building higher level semantic descriptors [3].

Low-level descriptors can be extracted “ad hoc” from the considered signal, independently from the adopted coding scheme, which requires therefore additional computation or, more efficiently, they can be calculated form the intermediate data representation available during the encoding process, as described in [13]. Usually the effectiveness of the low-level descriptors is evaluated in terms of retrieval capabilities, measured by estimating the Recall and Precision (RP) indicators which can also be used as objective function in the optimal design of descriptors and the associated metrics. The coded content and its descriptions can be packaged in a unique stream by using standard tools such as, for example, those provided by ISO MPEG-21 [5] or by SMPTE Material eXchange Format [19]. Besides providing interoperability, the above tools ensure a consistent link between the content and its description.

To jointly consider the problem of content coding and content analysis, i.e., to produce a unique compressed stream embedding the content and description information, could provide anyway additional advantages, namely:

- Descriptors are freely and immediately available to the decoder and there is no need to reprocess the content to extract them.
- The content and its description are in the same data flow which avoid the need to use additional tools such as for example MPEG-21 [5] to create a consistent link between the content and its description.
- The total rate, compressed bitstream plus description, should decrease, by removing possible information redundancies.

The concepts underlying this problem have been formulated in a seminal paper by R.W. Picard [15] by including in the overall coding objectives also the “Content access work” (in the cited paper, the so called “Fourth criterion”) added to the three classic criterion, “Bitrate”, “Distortion”, “Computational cost”. This basic idea was then further developed in [9], where an image codec providing an easy access to the spatial image content was proposed and evaluated. Examples of other developpements of this idea can be found in [18, 16], where an image coding method explicitly designed to ease the task of content access has been introduced. Specifically, in [16] the proposed coding method uses quite sophisticated approaches such as Segmentation-Based Image Coding, Vector Quantization, Coloured Pattern Appearance Model, and leads the possibility to access image content descriptors with reduced computational complexity. In Schaefer et al. [18], the proposed image encoder is based on CVPIC technique (Color Visual Pattern Image Coding). The data available at the output of this encoder, for each block in which the image is split, describe the color (in CIEL*a*b* space) and the planarity or the irregularity (presence of edges, corners, gradient). Then, the image descriptor is represented by edge map and color map. Starting from the proposed results, this technique appears to offer good retrieval performance, but it is not possible to evaluate the performance of encoding. Zhang et al. [24] proposed a compression method for videos, where from each shot of the video the ”key-objects” are extracted. They are particular descriptors that can characterize the entire shot, and they are outlined in terms of color, texture, shape, motion and their life cycle. Similarly to the coding standard MPEG-4, the metadata extracted for all key-objects are used for building the coded stream. Authors do not provide any information about system performance in compression and retrieval work. In [20] a technique for coding collection of images is described. It allows retrieving of content information working directly inside the coded domain. Each image is decomposed in a group of objects associated to semantic indexes (like ”tree”, ”house”). Then, the different areas are split in rectangular blocks that are coded separately, using a JPEG-like coding method. Hence, the coded stream is made by indexes, by their spatial relations and by the ”true” compressed content. Other works, such as that presented in [12], try to reduce the access work needed to generate a content description, by adopting the ”Rough Indexing” paradigm. Following this approach, low resolution/level descriptors are extracted requiring only a partial decoding. Following this directions, in [1] the case of scalable joint data and descriptor encoding of image collections have been considered.

In this paper we propose an extension of the previously described ideas to video sequences. The proposed technique is based on the appropriate combination of Vector Quantization (VQ) and wavelet based scalable video coding (SVC). More specifically, the key-pictures of each video GOP are encoded at a first draft level by using an optimal visual-codebook, while the residual errors are encoded using a JPEG2000 approach. The same visual-codebook is also used to encode all the key-pictures of a given video shot where boundaries are dynamically estimated. In this way, the visual-codebook is freely available as an efficient visual descriptor of the considered video shot, and also a rough temporal segmentation is implicitly provided. The paper is organized as follows. In Section 2, the concepts of content description and midstream information are highlighted. Section 3 presents the scalable video coding system, while Section 4 provides its experimental evaluation and possible uses of the proposed framework. Concluding remarks are given in the final section.

2 Content descriptions and Midstream Information

According to MPEG-7 standard [11], a content description is formed by a set of instantiated Descriptions Schemes (DS) and their associated Descriptors (D). Broadly speaking, we can say that Descriptors represent metadata extracted from a given content, at different semantic level, while Description Schemes are used to specify the spatio-temporal and logical/semantic relationship among Descriptors, and to associate them to the described content.

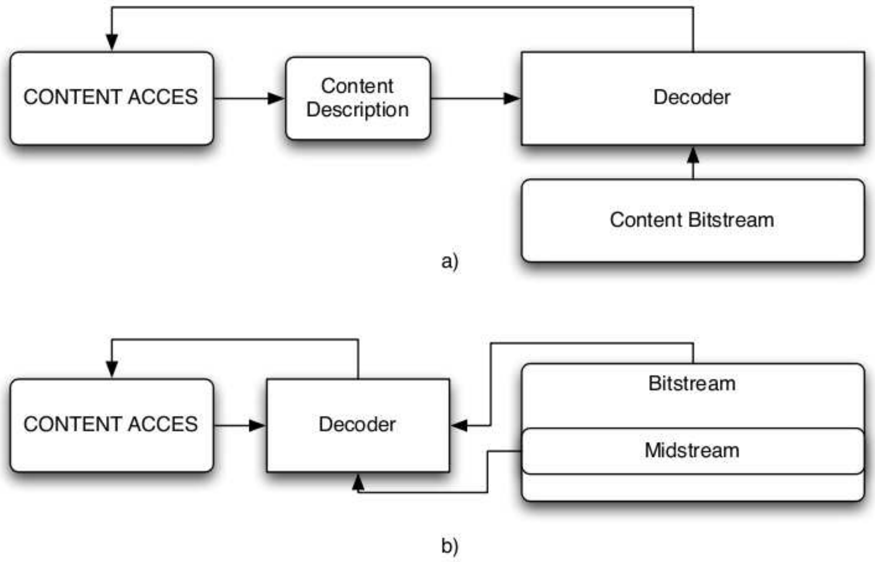

Fig. 1 Content Access: a) content description based access; b) midstream based access.

Content descriptions are mainly intended to facilitate the task of content browsing and retrieval which can be performed following the logical operation chain described in Fig.1.a. For example, if we want to retrieve all the scenes in a video containing red colors, a query can be performed through the content description and successively the retrieved content can be accessed by decoding the corresponding part of the bitstream.

A different scenario is described in Fig.1.b, where the content description is encapsulated in the bitstream, specifically in the midstream. Assuming that the costs of accessing content description and midstream information are comparable, the previous goal of scene retrieval can be achieved by searching for a specific content through the midstream information. Then, since the midstream is an integral part of the content bitstream, only an additional partial decoding will be required to fully access the retrieved content, improving therefore the overall efficiency.

The midstream construction can be performed, at least in principle, following different approaches. Assuming, for example, that a coding method and a content description are given, an embedding procedure could determine a bilateral correspondence between description elements and bitstream elements. This operation is clearly difficult, and it is not guaranteed that a suitable transformation could always be found. While this approach provides backward compatibility with existing coding methods, it requires a reverse transformation prior decoding, needed to recover the proper compressed bitstream format. A more convenient approach requires to jointly define both the coding method and the content descriptors. This is somehow similar to what has been done in the context of MPEG-4 standard concerning the Media Object Based coding part. The main difference is that while in MPEG-4 the identification of objects in video frames was mainly intended as a way to improve the compression performance, in the present work the main objective is twofold: to add a midstream layer containing meaningful content descriptors, and to optimize rate distortion coding performance.

2.1 Scalable bitstream and content descriptions

The scalable approach to content coding is the natural context to embed also scalable descriptions. In image/video scalable coding, the bitstream is mainly organized into different layers which fragment the whole content along the scalability axes, as described in Fig.2.a, where, to simplify the explanation, only spatial and quality scalability have been considered. The full quality for HD spatial resolution is recovered by decoding only the part labeled with HD. Adding the enhancement information (Full HD) allows to extract also the higher spatial resolution. If the stream is cut in a specific point, it is also possible to decode the signal

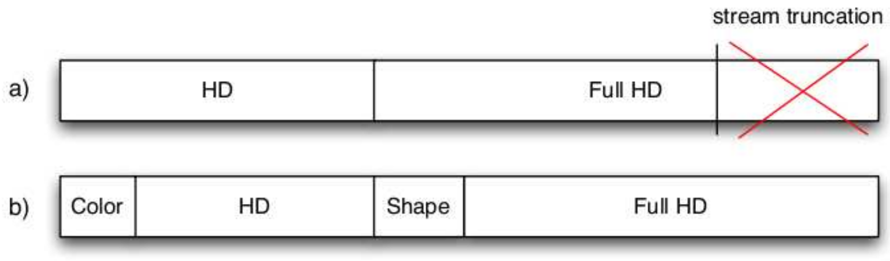

Fig. 2 a) Spatial and quality scalable bitstream; b) Spatial and quality scalable bitstream and midstream.

at a lower quality. Content Descriptors can be potentially embedded in each layer of the bitstream producing at least a spatially scalable description. For example, the lower spatial resolution, which corresponds (see Fig.2.b) to a low resolution version of the encoded content, could embed color and texture information, while the enhancement layers, which correspond to details information at higher resolution, could embed shape/contour descriptors.

A scalable bitstream with embedded descriptors (midstream) can therefore be scaled by an extractor operating on the content and/or on the descriptions, as exemplified in Fig. 3.

The scaled bitstream can then be decoded considering the content and/or the description at the desired resolution. In this work we have considered only low-level descriptors. This choice was guided by two reasons. First, the semantic level of the information related to low-level descriptors is closer to that carried out by the conventional compressed bitstream elements, and therefore it is should be feasible to easily embed them in the bitstream. Second, while low-level descriptors can be successfully used to support the generation of high level semantic descriptions, the latter information is often represented as a textual metadata (e.g., MPEG-7 events) and then it would be hardly encapsulated in the midstream without increasing the overall bit-rate. Another relevant aspect to be taken into account is the increase of complexity at the decoder side. When a piece of content is encoded (compressed)

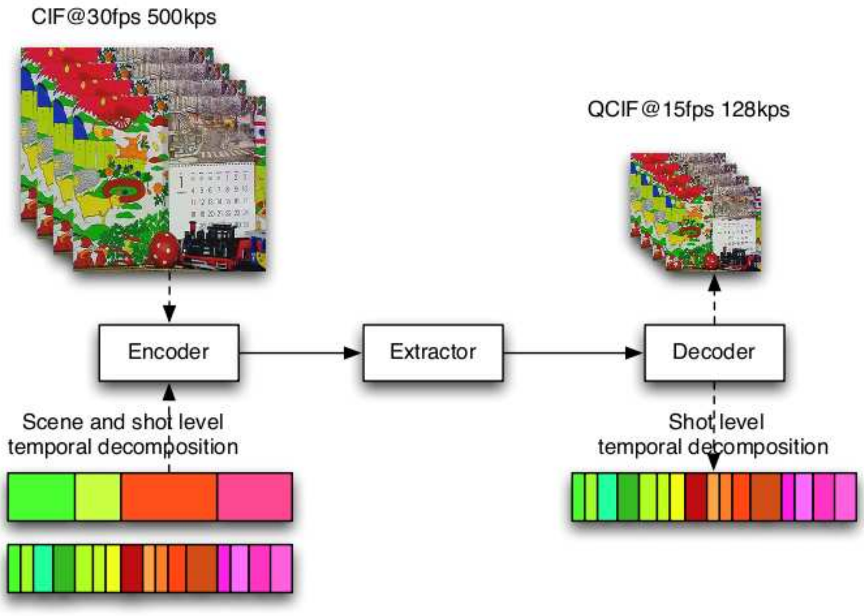

Fig. 3 Scalable coding and description rationale.

as a descriptor, a certain amount of operations needs to be performed. While this augmented complexity could not be a big deal at the encoder side, it is relevant for a decoder which usually has to be designed as simple as possible in order to reduce the amount of needed resources, including power consumption.

3 Scalable Video Coding and description embedding

As an example of the concepts presented in the previous sections, a scalable video codec which embeds low-level visual descriptors is hereafter proposed. Although pure Vector Quantization (VQ) based coding method have been in some way surpassed, in terms of coding performance, by hybrid coding methods (Motion Compensation with DCT/DWT), the code-book provided by VQ methods have been recognized as a good low-level descriptor of visual content [17]. Starting from this consideration, a new hybrid video codec architecture has been designed which properly combines the use of vector quantization and classic hybrid video coding.

In conventional video coding, the sequence of frames is initially subdivided into Group of Pictures (GOP), as shown in Fig. 4. Each key-picture is then independently encoded in Intra mode, while B and P pictures are encoded by exploiting temporal redundancy. Please note that hereafter the word key-picture is used to indicate a frame, belonging to the video sequence, encoded by using only intra mode (I frame), and it does not in general correspond to the concept of key-frame used in video indexing. Additionally, Fig. 4 specifically refers to a Hierarchical B-predicted Picture temporal decomposition defined in H.264/MPEG4-AVC coding standard, which provides dyadic temporal decomposition.

With respect to a conventional video codec, the method proposed in this paper uses a different approach to key-picture encoding. More specifically, as it can be seen in Fig.5, all key-pictures belonging to a given visually coherent sequence of GOPs, indicated as video shot, are encoded by using a vector quantization based predictive scheme, while B frames of a given GOP are encoded by applying a conventional motion estimation and compensation method. The coding process starts by considering the first GOP (e.g., GOP1 in Fig. 4) which also represents the beginning of the first video shot. Its key-picture is selected as a shot representative key-frame and decomposed into rectangular blocks which are then transformed into visual-vectors and used to build a visual-codebook, by applying the VQ training procedure described below. The key-frame and all the key-pictures of the GOPs belonging to the first video shot are then vector quantized by using the previously estimated visual-codebook, while the residual signals are generated by subtracting the quantized frames form the corresponding originals, as described in Fig.5. The visual-codebook, the key-pictures indexes and residuals are then entropy coded and properly placed in the bitstream. The shot boundary detection is performed by dynamically evaluating the VQ prediction error. When

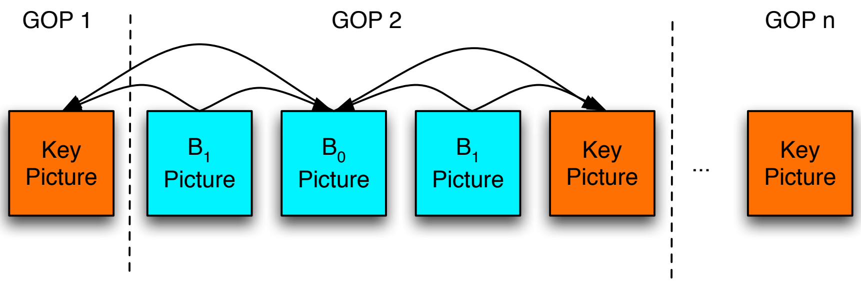

Fig. 4 Temporal decomposition. Note: the first GOP of a video sequence is composed only by one I frame.

the vector quantization of a new key-picture (next GOP) produces a reconstruction error higher than a predefined threshold, the used visual-codebook can not represent anymore the incoming visual content with sufficient precision. Consequently a shot-cut is declared, and a new visual-codebook is estimated by using the code-vectors extracted form the new key-picture, which became the key-frame of the new shot, as training data. It is worth to mention that the obtained video shots only broadly correspond to single camera record mainly due to the fact that shot-cut detection is performed only on key-pictures. Nevertheless the detected boundaries provide a good indication of visually coherent temporal segments.

3.1 Joint optimal visual-codebook design

In the following it is assumed that the Visual Codebook $VC(l,k)$ indicates a set composed by k code-vectors of l^2 elements, where l corresponds to the size of square image blocks. According to the proposed coding scheme, the visual-codebook is used both to exploit redundancy and, for example, to characterize visual patterns (image blocks) extracted from the considered video shot. Consequently the selection of the best value of l and k can not be based only on the evaluation of coding performances but also taking into account content

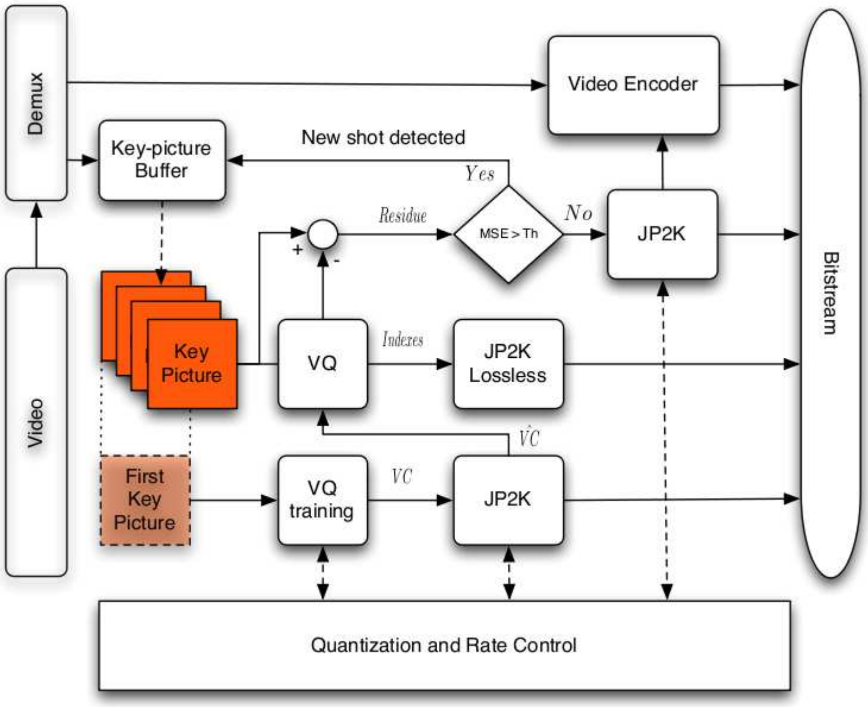

Fig. 5 The proposed video encoder.

characterizzaztion. This goal can be achieved by finding the pair $(\bar{l},\bar{k})$ which minimize the following cost function:

$$J(l,k) = VQ_{cost}(l,k) + PR_{cost}(l,k) \tag{1}$$

where $VQ_{cost}(l,k)$ is the cost function associated to the vector quantization process of the key-picture, and $PR_{cost}(l,k)$ expresses how the codebook $VC(l,k)$ allows for a good discrimination of the considered shot with respect to those already encoded. The first term of the function is given by the weighted sum $VQ_{cost}(l,k) = \alpha R_{VC(l,k)} + \beta D_{VC(l,k)}$ where $R_{VC(l,k)}$ and $D_{VC(l,k))}$ are the rate and the distortion associated to the vector quantization process. The second term of Equation (1) is given by $PR_{cost}(l,k) = \gamma(1 - Prec(l,k))$

where $Prec(l,k)$ is the precision of inter-shot classification, i.e., the percentage of correctly classified key-pictures with respect to the shots already encoded.

Clearly, the optimization of the cost function $J(\cdot,\cdot)$ is a computationally intensive task: it is needed to compute a RD function for several values of (l,k) and it is required to previously estimate its average behaviors of Precision and Recall by using a ground truth. While efficient ways to solve this problem are under investigation, (l,k) values can be selected in order to obtain the best compression or alternatively the best classification Precision (or alternatively Recall).

3.2 VC based dissimilarity

Traditionally employed for coding purposes [8], visual-codebooks (VC) have been successfully proposed in [17] as an effective low-level feature for video indexing. Assuming to quantize the visual content by using a VC, the dissimilarity between two images and/or two shots can be estimated by evaluating the distortion introduced if the role of two VCs is exchanged. More formally, let C_i, $i = 1, .., n$ be an image (or a shot), N_i the number of visual-vectors (VV) obtained after its decomposition into blocks, and let VC_j be a generic visual-codebook generated by a vector quantizer. The reconstruction error can then be measured by evaluating the average distortion $D_{VC_j}(S_i)$, defined as:

$$D_{VC_j}(S_i) = \frac{1}{N_i} \sum_{p=1}^{N_i} \|vv_i(p) - vc_j(q)\|^2 \ , \tag{2}$$

where $vc_j(q)$ is the codeword VC_j with the smallest euclidean distance from the visual-vector $vv_i(p)$, *i.e.*:

$$q = \arg\min_z \|vv_i(p) - vc_j(z)\|^2 \ . \tag{3}$$

Now, given two codebooks VC_h and VC_j, the value:

$$|D_{VC_h}(C_i) - D_{VC_j}(C_i)| \tag{4}$$

can be interpreted as the dissimilarity between the two codebooks, when applied to the same visual content C_i. A symmetric form, used in [17] to estimate the dissimilarity measure between different images C_i and C_j can, thus, be defined as:

$$\phi(C_i, C_j) = |D_{VC_j}(C_i) - D_{VC_i}(C_i)| +$$

$$+|D_{VC_i}(C_j) - D_{VC_j}(C_j)| \tag{5}$$

where VC_i and VC_j are in this case the optimal codebooks for the shot C_i and C_j, respectively. The smaller $\phi(.)$ is, the more similar the images (or shots) are. Note that the proposed dissimilarity measure is based on the cross-effect of the two codebooks on the two considered shots. In fact, it may be possible that the majority of blocks of one shot (for example C_i), can be very well represented by a subset of the codewords of the codebook VC_j. Therefore VC_j can represent C_i with a small average distortion, even if the visual-content of the two shots is only partly similar. On the other hand, it is possible that codebook VC_i does not lead to a small distortion when applied to C_j. So the cross-effect of codebooks on the shots can generally reduce the number of wrong classifications.

4 System evaluation

In this section the coding performance, in term of rate-distortion, an example of joint optimal visual-codebook design and the access content functionalities, provided by the embedded content midstream are presented and discussed. For all proposed experiments, the algorithm used to train the vector quantizer is the accelerated *k-means* extensively described in [7]. This algorithm has been preferred to the split LBG because of its efficiency. Nevertheless future extension of this work will consider the use of variable VC length. Given the best

VC, all codewords are arranged in order to represent the VC as an image. Basically this task is the dual process used to generate the visual-vector set from an input image. The VC image is then encoded, lossless or lossy, by using JP2K. The compressed bitstream is sent to the multiplexer, while the decoded version of this signal ($\hat{VC}$) is passed to the next coding stage as described in Fig. 5. The indexes of vector quantized images are then coded by using lossless JP2K while the residual images are compressed with lossy JP2K.

4.1 Coding effectiveness

Figure 6 reports the average Rate-Distortion curves, obtained for a video composed by the concatenation of four coding test sequences, namely, Mobile Calendar, City, Crew and Harbor. As it can be seen, the RD values provided by the proposed encoder are lower than that obtained using the pure JP2K intra coding method but anyway higher than the values given by the JPEG method. The cause of this loss are the artificial high frequencies introduced by the blockiness that characterize the quantized image (predictor).

4.2 Coding efficiency and shot-cut detection

In the previous section it has been highlighted how the selection of the most suitable parameters (l, k) can be performed by minimizing the cost function $J(\cdot, \cdot)$. Beside content description provided by the visual-codebook itself, an additional content information is given by the knowledge of shot boundaries. The shot-cut detection, indirectly performed by the system, is mainly influenced by the threshold value Th (see Fig. 5). In fact when the normalized difference between the prediction error of the incoming key-picture and the prediction error of the last encoded key-picture, is greater Th, a shot-cut is declared, and consequently

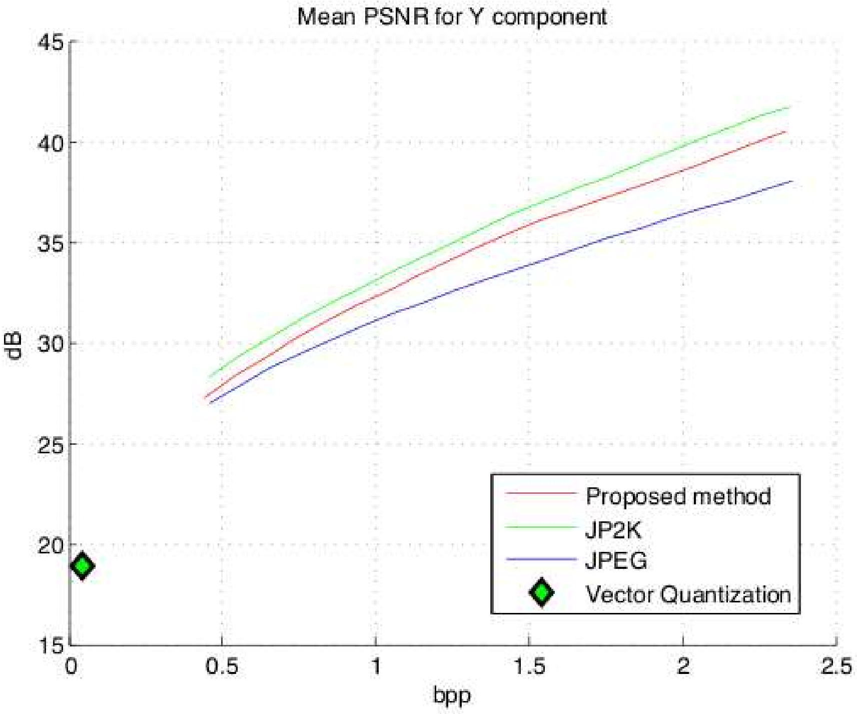

Fig. 6 Rate-Distortion curve using JP2K and the proposed VQ based encoding ($l = 16$ and $k = 8$).

a new visual-codebook is estimated. To evaluate the impact of this threshold on the system performance, the proposed method has been used to encode the video test set specified in TRECVID shot boundary detection task [14]. To this aim, different behaviors have been analyzed: the shot-cut Precision/Recall curve and the coding efficiency, obtained varying the threshold Th. As expected, the shot-cut detection performance reported in Fig. 8 are not comparable with those provided by state-of-the-art algorithms, mainly because the proposed system can not properly detect gradual transitions. Better results have been obtained for what concerns coding efficiency. As it can be seen in Fig. 7, for a quality higher than $32dB$, the threshold value does not substantially affect coding efficiency. This allows the selection of Th in order to obtain the desired value of shot-cut Precision or Recall according to expected temporal segmentation objective.

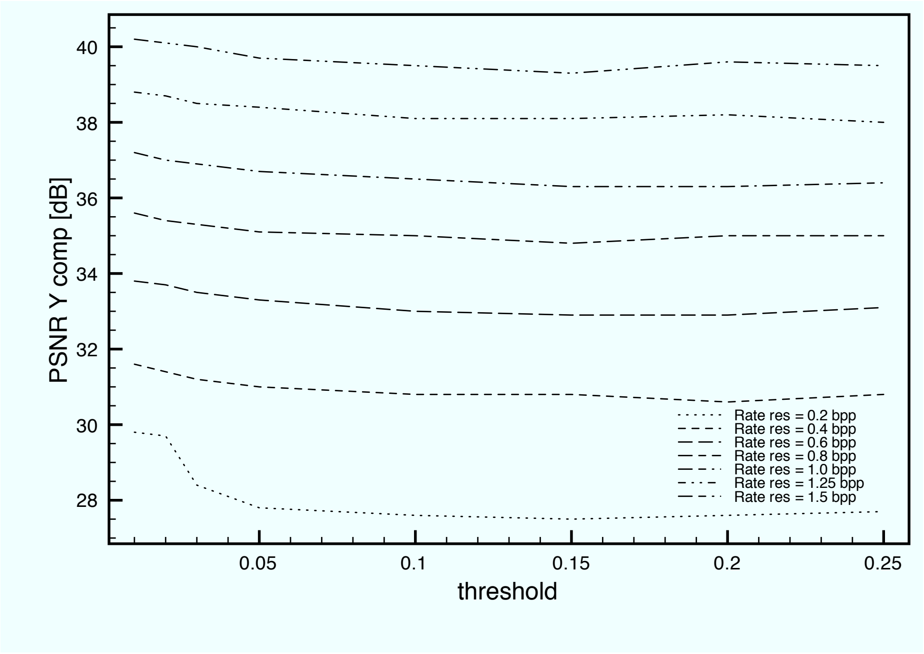

Fig. 7 Average Rate-Distortion curves for different values of threshold Th ($l = 16$ and $k = 16$).

4.3 Video Browsing

In Fig. 9 it is shown an example of the different levels of quality at which the video shot can be browsed. The lowest level corresponds to the representation of a shot only by using the associated visual-codebook. It is interesting to note that this minimal representation gives a quick idea of the color patterns that characterize the visual content. It is also possible to visualize a video through the vector quantized version of the first key-picture (second and third columns in Fig. 9) or in more detailed views according to the available quality layers

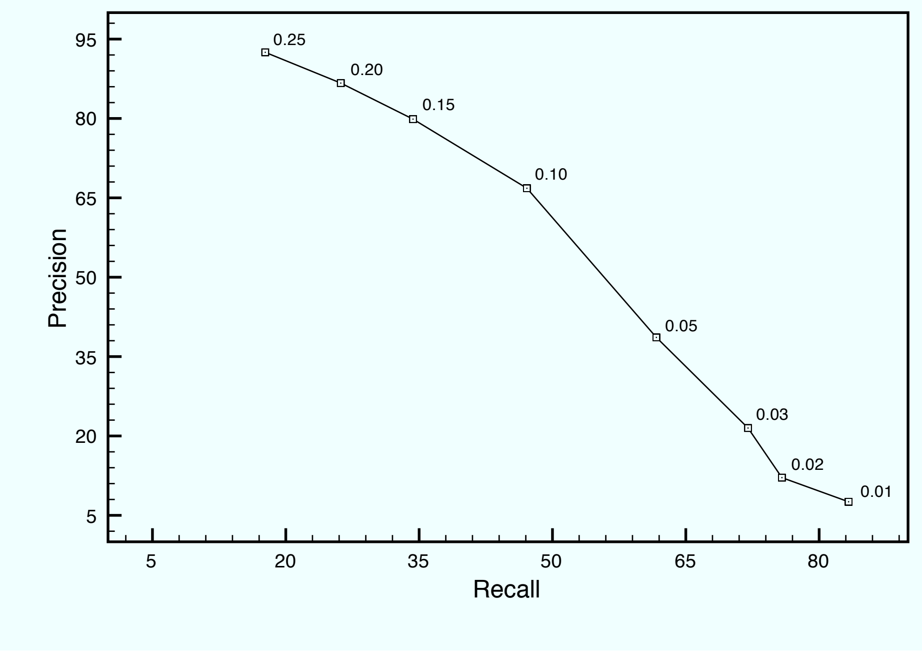

Fig. 8 Average Shot-cut Precision and Recall curve for different values of threshold Th ($l = 16$ and $k = 16$).

in the compressed bitstream.

4.4 Video Shot clustering

Beyond coding and fast browsing, the VC can be used to efficiently cluster video shots having similar visual content according to the measure presented in Section 3.2. The results, obtained by applying the above procedure to a movie trailer, are shown in Fig. 10, where each shot in a given cluster has been represented by its key-frame. In general shots belonging to a given cluster do not necessarily share a common underlying semantic as intended by humans. However if the clustering is considered as the first step of a more complex process, such as, for example, the extraction of hierarchical summaries or Logical Story Units as

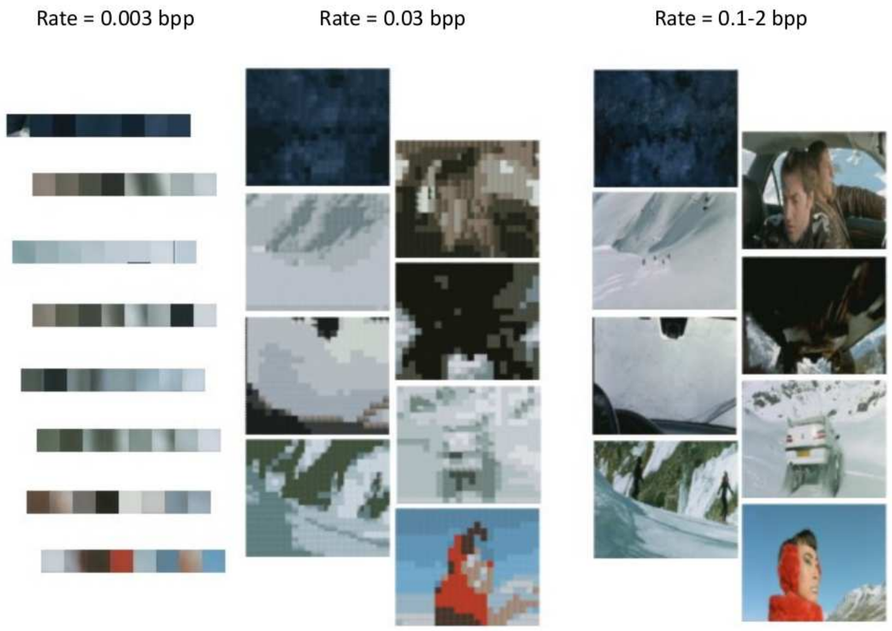

Fig. 9 Examples of possible browsing resolution levels and the associated rates.

proposed in [3], the effectiveness of the visual-codebook from the point of view of content description becomes more evident.

5 Conclusion

This paper proposes an efficient method for scalable video coding with embedded low-level descriptors. The main contribution concerns the ability of encasulating in the compressed bitstream also some low-level descriptors allowing for a new scalability dimension. It has been shown that this embedding can be performed with relatively small RD performance loss, and how this low-level descriptors could be used in content based browsing and clustering applications. Although the presented results are in some way limited, they adequately

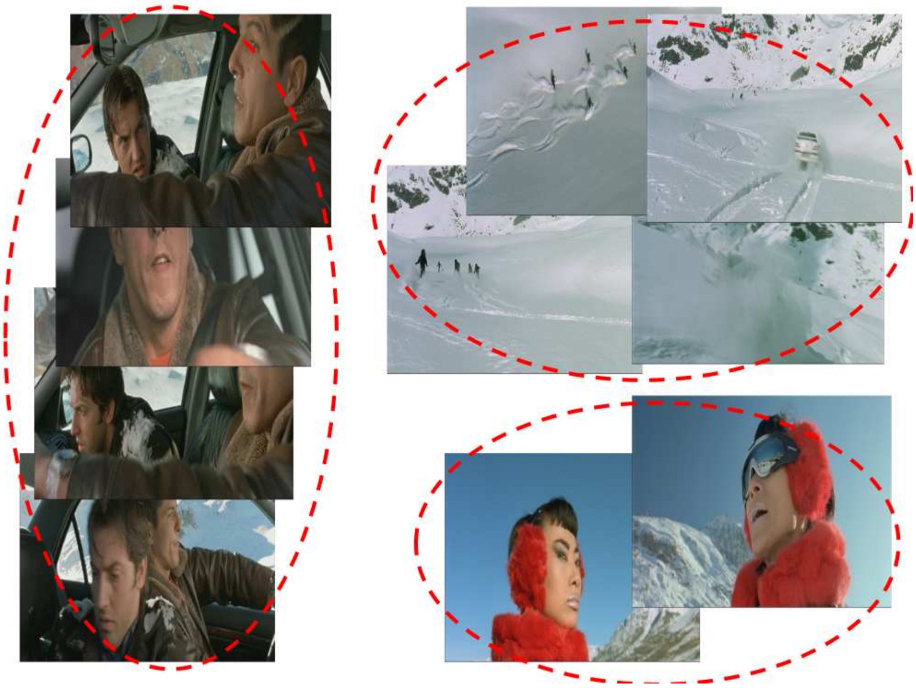

Fig. 10 Shot clustering based on visual-codebook distances.

support the idea of jointly perform video coding and content analysis, producing a compressed bitstream embedding also relevant information for content based description.

References

1. Adami, N., Boschetti, A., Leonardi, R., Migliorati, P.: Scalable coding of image collections with embedded descriptors. In: Proc. of MMSP-2008, pp. 388–392 (Cairns, Queensland, Australia, October 2008)
2. Adami, N., Signoroni, A., Leonardi, R.: State-of-the-art and trends in scalable video compression with wavelet-based approaches. IEEE Trans. Circuits and Syst. Video Technol. **9**(17), 1238–1255 (2007)
3. Benini, S., Bianchetti, A., Leonardi, R., Migliorati, P.: Extraction of significant video summaries by dendrogram analysis. In: Proc. of international Conference on Image Processing ICIP’06. Atlanta, GA, USA (2006)
4. Burnett, I.S., Pereira, F., de Walle, R.V., Koenen, R.: The MPEG-21 Book. John Wiley & Sons (2006)

5. Chang, S.F., Ma, W.Y., A., S.: Recent advances and challenges of semantic image/video search. In: Proc. of ICASSP-2007 (Hawaii, USA, April 2007)

6. Elkan, C.: Using the triangle inequality to accelerate k-means. In: Proc. of ICML, pp. 147–153 (Washington DC, 2003)

7. Gersho, A., Gray, R.M.: Vector quantization and signal compression. Kluwer Academic Publishers, Norwell, MA, USA (1991)

8. Hanjalic, A., Lagendijk, R., Biemond, J.: Efficient image codec with reduced content access work. In: Proc. of ICIP, pp. 807–811 (Kobe, Japan, 1999)

9. Izquierdo, E., al.: State of the art in content-based analysis, indexing and retrieval. In: IST-2001-32795 SCHEMA Del. 2.1, Feb. 2005 (2005)

10. Manjunath, B., Salembier, P., Sikora, T.: Introduction to MPEG-7: Multimedia Content Description Language. John Wiley & Sons (2002)

11. Morand, C., Benois-Pineau, J., Domenger, J.: Scalable indexing of HD video. In: Proc. Content-Based Multimedia Indexing 2008, pp. 417–424 (2008). DOI 10.1109/CBMI.2008.4564977

12. Morand, C., Benois-Pineau, J., Domenger, J.P., Mansencal, B.: Object-based indexing of compressed video content: From sd to hd video. In: ICIAPW ’07: Proceedings of the 14th International Conference of Image Analysis and Processing - Workshops, pp. 71–76. IEEE Computer Society, Washington, DC, USA (2007). DOI http://dx.doi.org/10.1109/ICIAPW.2007.34

13. NIST: Guidelines for the trecvid 2007 evaluation - shot boundary detection task (2008). Http://www-nlpir.nist.gov/projects/tv2007/tv2007.html

14. Picard, R.W.: Content access for image/video coding: The fourth criterion. Tech. Rep. 295, MIT Media Laboratory - Perceptual Computing Section, Cambridge, USA (1994)

15. Qiu, G.: Embedded colour image coding for content-based retrieval. Journal of Visual Communication and Image Representation **15**(4), 507–521 (2004)

16. Saraceno, C., Leonardi, R.: Indexing audio-visual databases through a joint audio and video processing. International Journal of Imaging Systems and Technology **9**(5), 320–331 (1998)

17. Schaefer, G., Qiu, G.: Midstream content access of visual pattern coded imagery. In: Proc. of 2004 Conference on Computer Vision and Pattern Recognition, pp. 144–149 (June 2004)

18. Standard, S.: Material exchange format (mxf) - file format specification, smpte 0377-1-2009 (2009)

19. Swanson, M.D., Hosur, S., Tewfik, A.H., Ansari, R., Smith, M.J.T.: Image coding for content-based retrieval. In: Visual Communications and Image Processing ’96, vol. 2727, pp. 4–15. SPIE, Orlando, FL, USA (1996). URL http://link.aip.org/link/?PSI/2727/4/1
20. Taubman, D.: High performance scalable image compression with EBCOT. IEEE Trans. Image Processing **9**, 1158–1170 (2000)
21. Taubman, D.S., Marcellin, M.W.: JPEG 2000: Image Compression Fundamentals, Standards and Practice. Kluwer Academic Publishers, Norwell, MA, USA (2001)
22. Wang, H., Cheung, N.M., Ortega, A.: A framework for adaptive scalable video coding using wyner-ziv techniques. EURASIP Journal on Applied Signal Processing **Article ID 60971**(doi:10.1155/ASP/2006/60971) (2006)
23. Zhang, H., Wang, J., Altunbasak, Y.: Content-based video retrieval and compression: a unified solution. In: Image Processing, 1997. Proceedings., International Conference on, vol. 1, pp. 13–16 vol.1 (1997). DOI 10.1109/ICIP.1997.647372