Embedded Indexing in Scalable Video Coding

Nicola Adami, Alberto Boschetti, Riccardo Leonardi, Pierangelo Migliorati
DEA - University of Brescia, Signals and Communications lab.
Via Branze, 38, 25123, Brescia, Italy
Email:{name.surname}@ing.unibs.it

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 bit-stream by jointly optimizing content encoding and indexing performances. This approach provides a new type of bit-stream 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 (VQ) and Scalable Video Coding has been developed and evaluated. 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 conventional approach.The same visual-codebook is also used to encode all the key-pictures of a video shot, which boundaries are dynamically 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. 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 bitrate. An evolution of these coding schemes tries to obtain the scalability of the coded bit-stream without affecting too much the coding performance in terms of RD [15], [17], [2]. Scalable Image Coding (SC) methods, such as for example JPEG2000 (JP2K) [16], generate a bit-stream with a unique feature, 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 [9], [5], [10] which, for example, can be used in query by example operation or as a base for building higher level semantic descriptors [3]. Usually low-level descriptors are extracted "ad hoc" from the considered signal, independently from the adopted coding scheme, which requires therefore additional computation at the expense of a more complex end-user system. The effectiveness of the low-level descriptors in terms of retrieval capabilities is usually measured considering the Recall and Precision (RP) indicators which can the be used as objective function in the optimal design of descriptors and the associated metrics.

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 provides 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 a in the same data

flow which avoid the need to use additional tools such as for example MPEG-21 [4] to create al consistent link between the content and its description.

- The total rate, compressed bit-stream 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 [11] 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, "Bit-rate", "Distortion", "Computational cost". This basic idea was then further developed in [8], where an image Codec providing an easy access to the spatial image content was proposed and evaluated. Another example of application of this idea can be found in [14, 12], where an image coding method explicitly designed to ease the task of content access has been introduced. Specifically in [12] 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. Following this direction, 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 previous works to video sequences. The proposed technique is based on the appropriate combination of Vector Quantization (VQ) and wavelet based scalable video coding. 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 which 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.

2. Content descriptions and Midstream Information

According to MPEG-7 standard [10], 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.

Content descriptions are mainly intended to ease content browsing and retrieval which can be performed following the logical operation chain described Fig.1.a. For example,

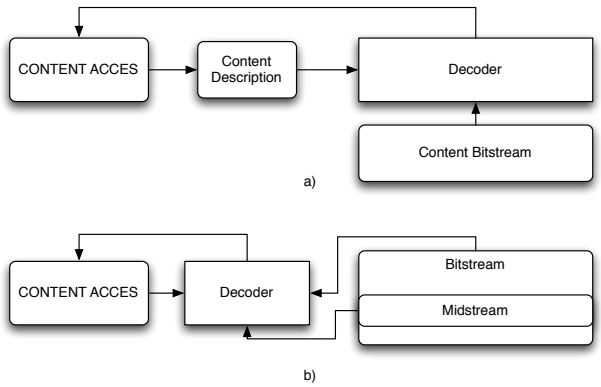

Figure 1. Content Access: a) content description based access; b) midstream based access.

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 bit-stream.

A different scenario is described in Fig.1.b, where the content description is encapsulated in the bit-stream, specifically in the midstream. Assuming that the cost of accessing to content description and to midstream information is comparable, the previous goal of scene retrieval can be achieved by searching for a specific content through the midstream information as previously described. Then, since the midstream is an integral part of the content bit-stream, only a 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 bit-stream 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 bit-stream format. A more convenient approach requires to jointly define both the coding method and the content descriptors. This is some how similar to what have been done in the context of MPEG-4 standard for what concern the Media Object Based coding part. The main difference is that while in MPEG-4 the identification of objects in video frames where 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

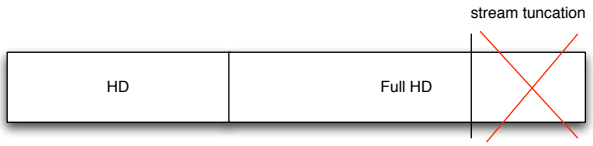

Figure 2. Spatial and quality scalable bit-stream.

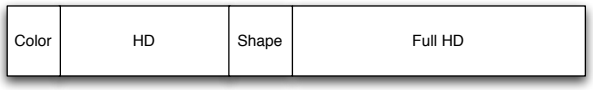

Figure 3. Spatial and quality scalable bit-stream and midstream.

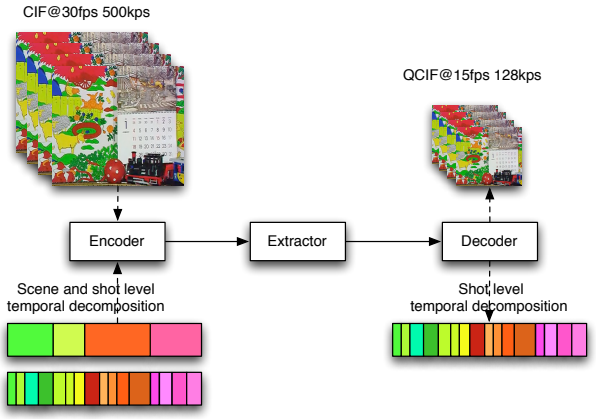

Figure 4. Scalable coding and description rationale.

optimize rate distortion performance.

Scalable bit-stream and content descriptions

The scalable approach to content coding is the natural context to embed also scalable descriptions. In image/video scalable coding, the bit-stream is mainly organized into different layers which fragment the whole content along the scalability axes, as described in Fig. 3 where, to simplify the explanation, only spatial and quality scalability have been considered. The full quality HD resolution of the signal is recovered decoding only the part labeled with HD. Adding the enhancement information (Full HD) allow to extract also the higher resolution. If the stream is cut in a specific point, it is also possible to decode the signal at a lower quality. Descriptors can be potentially embedded in each layer of the bit-stream producing at least a spatially scalable description. For example (see Fig.4) the lower spatial resolution, which correspond 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 bit-stream with embedded descriptors (midstream) can therefore be scaled by an extractor operating on the content and/or the descriptions, as reported in Fig 4.

The scaled bit-stream can then be decoded considering the content and/or the description at the desired resolution. In this work only low-level descriptors have been considered. This choice was guided by two reasons. First, the semantic of the information related to low-level descriptors is closer to that carried out by the conventional compressed bit-stream elements, and therefore it is quite easy its embedding in the bit-stream. Second, low-level descriptors can be successfully used to generate higher semantic level descriptions which would be hardly encapsulated in the midstream without increasing of 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) as a descriptor, a certain amount of operation need 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 taken as simple as possible in order to reduce the amount of computational 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 here presented. 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 and DCT/wavelet), the code-book provided by VQ methods have been recognized as good low-level descriptor of visual content [13]. Starting from this consideration, a new hybrid video codec has been designed which properly combines the use of vector quantization and classic hybrid video coding.

In conventional video encoder, the sequences of frames are initially subdivided into Group of Pictures (GOP), as shown in Fig. 5. Each key-picture is then independently encoded in Intra mode, while B and P pictures are encoded by exploiting temporal redundancy. Fig. 5 specifically refers to a Hierarchical B-predicted Picture decomposition which provides dyadic temporal decomposition.

Whit respect to a conventional video codec, the method

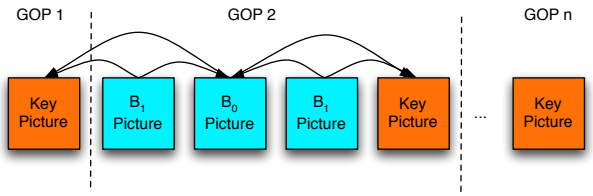

Figure 5. Temporal decomposition.

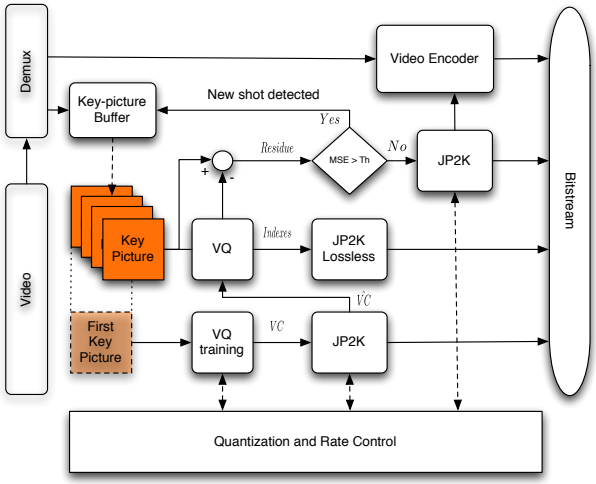

Figure 6. The proposed video encoder.

proposed in this paper uses a different approach to key-picture encoding. More specifically, as it can be seen in Fig. 6, all the key-pictures belonging to a given shot are encoded by using a vector quantization based predictive scheme. In more detail, each key-picture is initially decomposed into rectangular blocks which are transformed into code-vectors. When a new shot begin, an optimal visual-codebook is generated for the blocks of the first key-picture. All the key-pictures of the GOP are then vector quantized by using the previously calculated visual-codebook, and the residual signals are generated by subtracting the quantized frames form the original ones. The visual code-book is then quantized and placed at the beginning of the shot. The shot boundary detection is then performed by dynamically evaluating the VQ error. When the quantization of a new key-picture produces a reconstruction error higher than a predefined threshold means that the used visual codebook can not well represent the incoming visual content. Consequently a shot-cut is declared, and a new visual-codebook is estimated. All the B frames of a given GOP are then encoded applying a conventional motion estimation and compensation method.

Vector Quantization Training

The algorithm used to train the vector quantizer is the accelerated *k-means* described in [6]. The input images are initially decomposed into a set of vectors which are used as training data. A certain number of Visual Codebooks $VC(l,k)$ is generated varying the block size l and the number of codewords k in order to jointly find the best set of image blocks predictor, and the best set of image block descriptors for the whole set of key-pictures. This can be achieved by finding the pair $(\bar{l},\bar{k})$ which minimize the following cost function:

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

where $VQ_{cost}(l,k)$ is the cost function associated to the vector quantization of all the key-pictures in the considered video shot, 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 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.

VC based similarity

Traditionally employed for coding purposes [7], Visual Codebooks have been successfully proposed in [13] as an effective low-level feature for video indexing. Assuming to quantize the visual content by using a VC, the similarity 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:

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

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

$$q = \arg\min_z \|vv_i(p) - vc_j(z)\|^2 \,. \qquad (3)$$

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

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

can be interpreted as the similarity between the two codebooks, when applied to the same visual content C_i. A symmetric form, used in [13] to estimate the similarity measure

between different images C_i and C_j can, thus, be defined as:

$$\phi(C_i, C_j) = |D_{VC_j}(C_i) - D_{VC_i}(C_i)| + \\ +|D_{VC_i}(C_j) - D_{VC_j}(C_j)| \quad (5)$$

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 similarity 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.

Joint Coding-Indexing optimization

Obviously, the optimization of the cost function $J(.)$ is a computationally intensive task, since it is needed to compute the RD function cost associated to the vector quantization for several values of (l, k). Concerning inter classification precision, it is required to previously estimate its average behaviors by using a ground truth. Examples of the $VQ_{cost}(l, k)$ and $PR_{cost}(l, k)$ function. For some combinations of l and k there are possible saturation effects and this is the case, at least one of the two member of equation 1 can be removed from the optimization procedure. Given the best VC, all codewords are arranged in order to store the VC in an encoded image. Basically this task is the opposite 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 stream is sent to the multiplexer, while the decoded version of this signal ($\hat{VC}$) is provided to the following coding stage.

4. System evaluation

In this section the coding performance, in term of rate-distortion, and the access content functionalities provided by the possibility given by the embedded content midstream are presented.

Coding effectiveness

Figure 7 reports the average Rate-Distortion curves, obtained for a video composed by the concatenation of 4 classical coding test sequences, Mobile Calendar, City, Crew and Harbor. As it can be seen, the RD values provided by the proposed encoder are lower than the pure JP2K intra coding method but anyway higher than the old JPEG method. The cause of this loss are the artificial high frequencies introduced by the blockiness that characterize the quantized image (predictor).

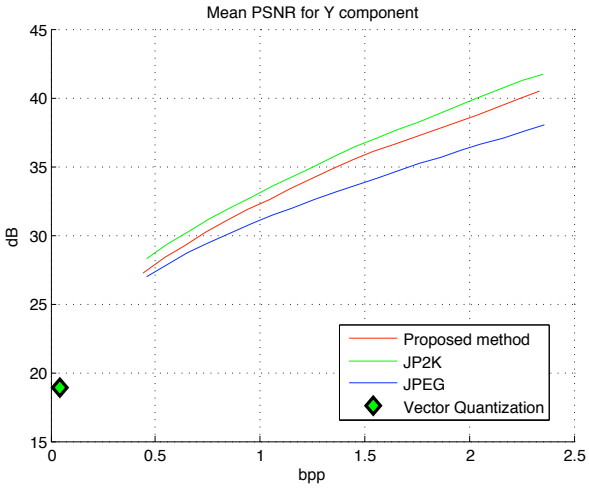

Figure 7. Y Rate-Distortion curve using JP2K and the proposed VQ based encoding ($l = 16$ and $k = 8$).

Video Browsing

In Fig. 8 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. Anyway, also this rough 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. 8) or in more detailed views according to the available quality layers in the compressed stream.

Video Shot clustering

Beyond coding and fast browsing, the VC can be used to efficiently cluster video shot having similar visual content. As and example, a shot clustering has been performed by using the similarity estimation measure, described in Section 3, which relies on the cross comparison of visual codebook. The results, obtained by applying the above procedure to a movie trailer, are shown in In Fig. 9 where each shot in a given cluster has been represented by the first key-picture.

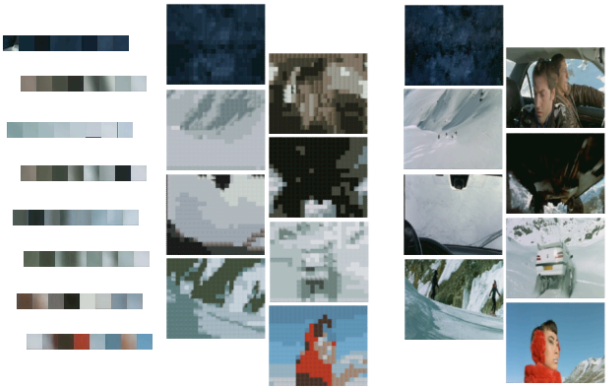

Figure 8. Examples of possible browsing resolution levels and the associated rates.

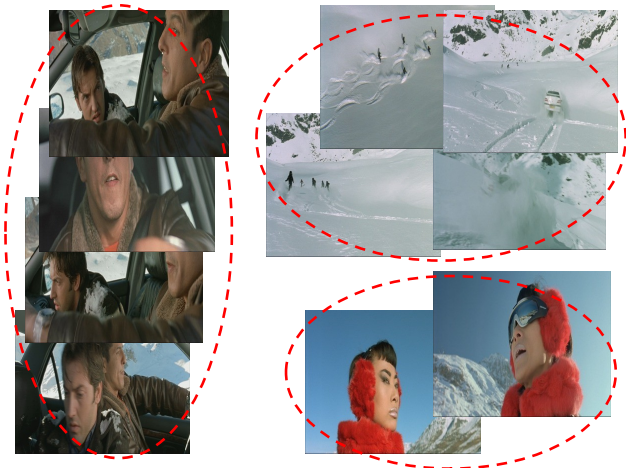

Figure 9. Shot clustering based on Visual-Codebook distances.

5. Conclusion

This paper proposes an efficient method for scalable video coding with embedded low-level visual descriptors. The main contribution concern the ability to embed in the compressed bit-stream also some low-level visual 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 effectively in content browsing and clustering applications. Although the presented results are in some way preliminary, they adequately demonstrate the idea of jointly perform content coding and content analysis, and producing a compressed bit-stream embedding also the content description.

References

[1] N. Adami, A. Boschetti, R. Leonardi, and P. Migliorati. Scalable coding of image collections with embedded descriptors. In *Proc. of MMSP-2008*, pages 388–392, Cairns, Queensland, Australia, October 2008.

[2] N. Adami, A. Signoroni, and R. Leonardi. 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] S. Benini, A. Bianchetti, R. Leonardi, and P. Migliorati. Extraction of significant video summaries by dendrogram analysis. In *Proceedings of international Conference on Image Processing ICIP'06*. Atlanta, GA, USA, 8-11 October 2006.

[4] I. S. Burnett, F. Pereira, R. V. de Walle, and R. Koenen. *The MPEG-21 Book*. John Wiley & Sons, 2006.

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

[6] C. Elkan. Using the triangle inequality to accelerate k-means. In *Proc. of ICML*, pages 147–153, Washington DC, 2003.

[7] A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[8] A. Hanjalic, R. Lagendijk, and J. Biemond. Efficient image codec with reduced content access work. In *Proc. of ICIP*, pages 807–811, Kobe, Japan, 1999.

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

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

[11] R. W. Picard. Content access for image/video coding: The fourth criterion. Technical Report 295, MIT Media Laboratory - Perceptual Computing Section, Cambridge, USA, 1994.

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

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

[14] G. Schaefer and G. Qiu. Midstream content access of visual pattern coded imagery. In *Proc. of 2004 Conference on Computer Vision and Pattern Recognition*, pages 144–149, June 2004.

[15] D. Taubman. High performance scalable image compression with EBCOT. *IEEE Trans. Image Processing*, 9:1158–1170, 2000.

[16] D. S. Taubman and M. W. Marcellin. *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, Norwell, MA, USA, 2001.

[17] H. Wang, N.-M. Cheung, and A. Ortega. 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.