Scalable Coding of Image Collections with Embedded Descriptors

N. Adami, A. Boschetti, R. Leonardi, P. Migliorati

Department of Electronic for Automation, University of Brescia
Via Branze, 38, Brescia, Italy

***Abstract*—With the increasing popularity of repositories of personal images, the problem of effective encoding and retrieval of similar image collections has become very important. In this paper we propose an efficient method for the joint scalable encoding of image-data and visual-descriptors, applied to collections of similar images. From the generated compressed bit stream, it is possible to extract and decode the visual information at different granularity levels, enabling the so called "Midstream Content Access". The proposed approach is based on the appropriate combination of Vector Quantization (VQ) and JPEG2000 image coding. Specifically, the images are encoded at a first draft level using an optimal visual-codebook, while the residual errors are encoded using a JPEG2000 approach. In this way, the codebook of the VQ is freely available as an efficient visual descriptor of the considered image collection. This scalable representation supports fast browsing and retrieval of image collections providing also a coding efficiency comparable with those of standard image coding methods.**

I. Introduction

Recently the problem of effective encoding and retrieval of similar image collections has become very important.

Traditional image 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 bit stream without affecting too much the coding performance in terms of RD. Scalable Coding (SC) methods, such as for example JPEG2000 (J2K) [1], 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 performances comparable with those of single point coding methods and requiring a very low sub-stream extraction complexity, actually comparable with read and write operations.

Considering the problem of content based analysis and retrieval of multimedia documents, a specific importance is related to the low-level descriptors of the considered audio-visual information. Usually these descriptors are extracted "ad hoc" from the considered signal, independently from the adopted coding scheme, which requires some more 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) performance. It could be an interesting approach to consider jointly the problem of data encoding and low-level descriptors representation in a unique data stream, optimized jointly with respect to the RD and the RP performance.

In this respect, in [2] the problem of 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", has been introduced as an innovative concept and an interesting new research field. This basic idea was then further developed in [3], 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 [4], [5], where an image coding method explicitly designed to ease the task of content access has been introduced. Specifically in [5] 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 this paper we have considered the specific case of scalable joint data and descriptor encoding of image collections. The proposed approach is based on a suitable combination of Vector Quantization and JPEG image coding. The basic idea is to represent a collection of similar images at different levels of resolution, using a progressive encoding method based on low-level visual descriptors, namely the Visual Codebook (VC) of a Vector Quantizer (VQ), used to generate a prediction of the image that will be refined in the next resolution levels. Both the visual descriptors and the prediction error are then encoded using a standard JPEG encoder. In this way, the visual descriptors are naturally embedded in the coded bit stream, without additional costs for their extraction. Moreover, the availability of the Visual Codebook allows for an efficient characterization of the image collection as a whole, very useful for both content-based retrieval and duplicate detection of image collections. The coding efficiency is also good, as demonstrated in the experimental evaluations.

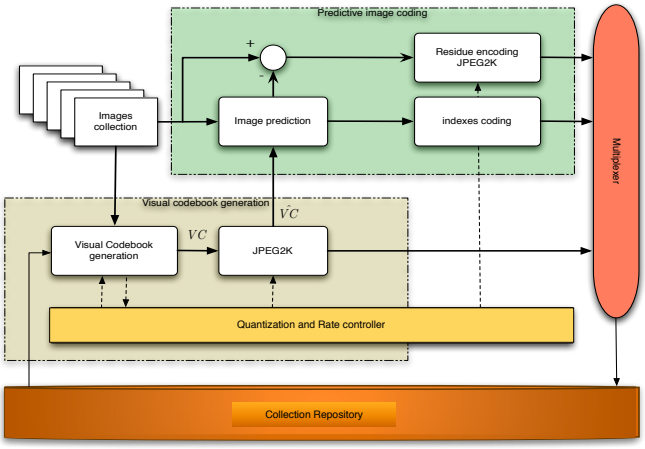

Fig. 1. The proposed hybrid encoder.

Fig. 2. High level bit stream structure.

II. The Proposed Hybrid Encoder for Collections of Similar Images

The hybrid codec proposed in this work, is shown in Fig. 1. Two main blocks can be identified: the optimal visual codebook estimator and the predictive encoder.

When a new group of similar images is added to the collection repository the coding process starts by finding the set of visual codebooks that provides the best RD tradeoff and, at the same time, allows to discriminate visually different groups of images. Once this set of visual codewords is found, all the images in the group are vector quantized generating a low quality version of the original content. In the second stage, the quantized images are used as a prediction for the corresponding input signal. The visual codebook and the reconstruction errors are then coded by using J2K, while the vector indexes are compressed by using a dedicated entropy coder. All this information is then packaged as shown in Fig. 2. This bit stream structure provides scalability at several levels of quality, depending on the way J2K is used. At least it is possible to extract and decode the images at two different qualities: the vector quantized version (base layer) of the original images, and the maximum coded quality obtained by adding all the residual information (enhancement layer). Moreover, if the enhancement layer is encoded by enabling quality scalability (in J2K) more working points can be obtained. By slightly modifying the proposed coding scheme it is also possible to get true spatial scalability. To achieve this goal, all the original images should be initially processed by applying the same wavelet decomposition used by the J2K encoder. Vector quantization and consequently prediction should then be applied to the low-pass spatial sub-band.

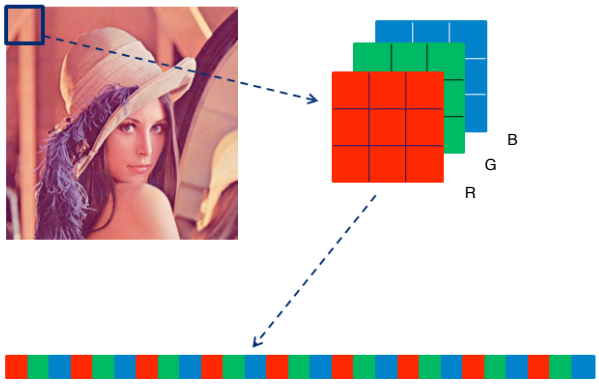

Fig. 3. Transformation of an image block into a vector assuming $l = 3$.

A. Generation of the optimal codebook

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, as shown in Fig. 3, 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 collection. This can be achieved by finding the pair $(\bar{l},\bar{k})$ which minimize the following cost function:

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

where $VQ_{cost}(l,k)$ is the cost function associated to the vector quantization of all the images in the considered collection, and $PR_{cost}(l,k)$ expresses how the codebook $VC(l,k)$ allows for a good discrimination of the considered images with respect to the collections already stored in the database. 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-collection classification, i.e., the percentage of correctly classified images with respect to the collections in the database.

B. Inter-collection classification

To enable image retrieval and/or classification a robust mechanism to evaluate image similarities is needed. Concern-

ing the choice of the low-level features used to represent the visual-content and the metric adopted to estimate similarities, a variety of methods can be found in the literature. However, even if some sophisticated techniques have been proposed, in most cases image indexes have been generated by means of color information, often simply relying on the color histogram in one of the available color spaces or on color spatial distribution. In this work, we describe each collection of similar images with a dictionary of visual codewords, obtained from a vector quantization process. Traditionally employed for coding purposes [7], visual codebooks have been successfully proposed in [8] 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 collections 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 collection), N_i the number of Visual Vectors (VV) obtained after its decomposition into blocks (see Fig. 3) 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 [8] to estimate the similarity measure between different images C_i and C_j can, thus, be defined as:

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

where VC_i and VC_j are in this case the optimal codebook for the image class C_i and C_j respectively. The smaller $\phi(.)$ is, the more similar the images are. Note that the similarity is based on the cross-effect of the two codebooks on the two considered image sets. In fact, it may be possible that the majority of blocks of one class (for example C_i), can be very well represented by a subset of codewords of 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 two classes can generally reduce the number of wrong classifications.

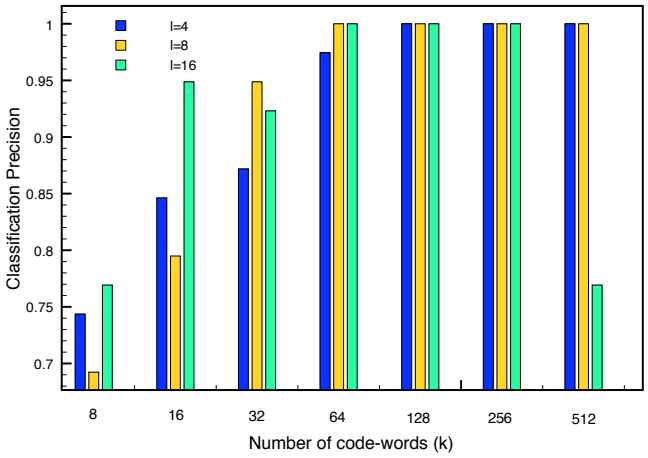

Fig. 4. Examples of inter collection classification precision curves.

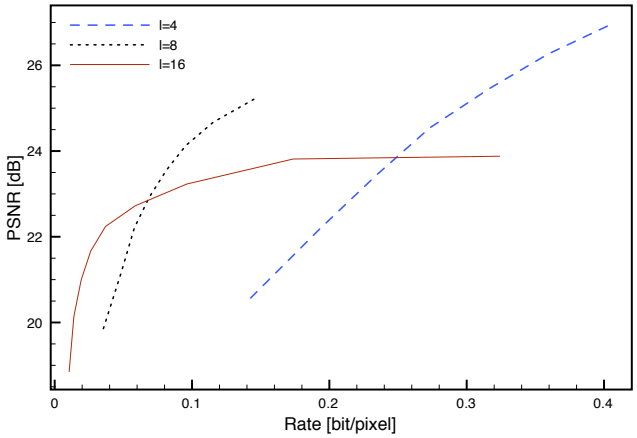

Fig. 5. Rate-Distortion curve obtained by applying only VQ (average value for a whole image collection).

C. *Selection and coding of the best visual codebook*

Obviously the optimization of the cost function $J(.)$ is a computationally intensive task, since it is needed to compute both the RD function cost associated to the vector quantization for several values (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 are reported in Figures 4 and 5 respectively. As it can be noticed, at some combinations of l and k there are possible saturation effects. If 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 its visual 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 (see Fig. 3). The VC image is then encoded, lossless or lossy, by using J2K. The compressed stream is sent to the multiplexer while the decoded version of this signal ($\hat{VC}$) is provided to the following coding stage.

D. Predictive image coding

Once the optimal visual codebook is generated and compressed, the scalable encoding of the whole collection is performed. Each image is initially quantized by applying the visual codebook $\hat{VC}$ (see Fig. 1). The generated visual codeword indexes are then entropy coded and sent to the stream multiplexer. The quantized image is then used as a prediction for the corresponding original input one. Finally the residual image, given by the difference between the input signal and its prediction, is encoded by using the J2K encoder.

III. System performance evaluation

In this section the coding performance and the browsing/retrieval capabilities provided by the proposed method are evaluated.

A. Coding effectiveness

The coding effectiveness of the proposed hybrid codec has been evaluated comparing its rate distortion curves with those achieved by separately encoding all the images with J2K. This has been accomplished by encoding several collections and averaging the results by using different sizes of the image blocks (l) and different numbers of visual codewords (k).

The best performances have been obtained, with $l = 16$ and few visual codewords $k \in [8, 64]$ (see Fig. 6), where the proposed hybrid encoder for rates ranging in $[0.5, 2.5]$ $bit/pixel$, or equivalently for quality values of $[35 : 40] dB$ behaves as well as his competitor. As shown in Fig. 7, decreasing the block size to $l = 8$ lowers also the performance for $k = 64$. When a small block size, e.g., $l = 4$ and/or a large number of visual codewords are used, the compression effectiveness significantly decreases. This behaviour can be explained by analysing the rate distortion curves associated to the pure vector quantization process and the characteristics of the residual error. In fact, the compression ratio provided by vector quantization systems are quite far from those provided by traditional image coding methods, especially at high rates. Moreover, as it has been previously shown, depending on the vector quantization algorithm used, saturation effects can occur (see Fig. 5). This implies that the prediction stage should use the lowest possible rate which provides the desired inter collection discrimination capabilities. Moreover, increasing the rate allocated to the vector quantization process increases the relative amount of high frequencies in the residual signal. Since J2K has been designed and tuned mainly to compress natural images, the performance of the residual image coding

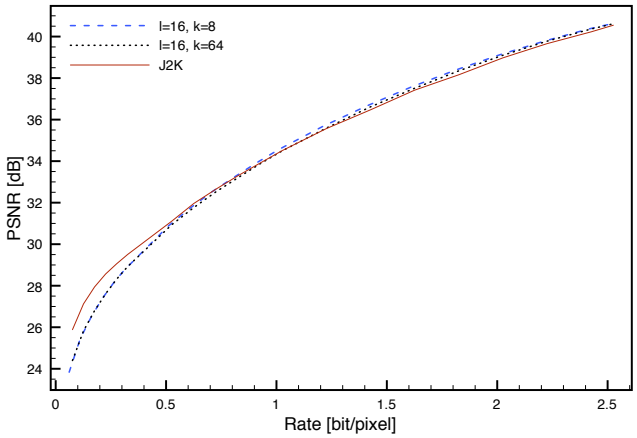

Fig. 6. Rate-Distortion curve comparison for $l = 16$ and $k = 8, 64$.

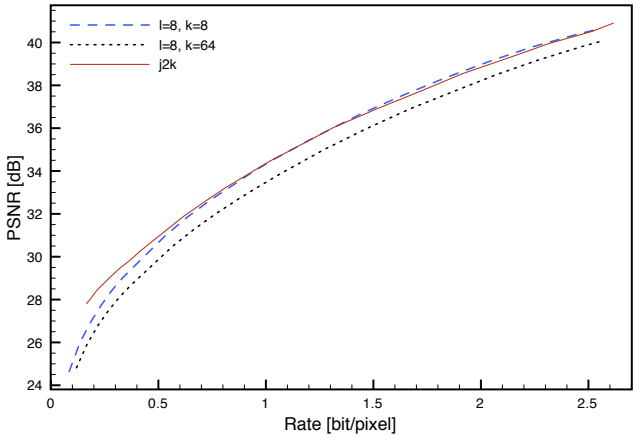

Fig. 7. Rate-Distortion curve comparison for $l = 8$ and $k = 8, 64$.

are significantly lowered. With respect to the use of pure J2K system, the complexity of the proposed method is clearly higher. However, this penalty is well balanced by the provided embedding of visual descriptors in the final bit stream, which can be used in image browsing and retrieval applications.

B. Scalable Browsing of Image Collection

In Fig. 8 it is shown an example of the different levels of details at which the collection repository can be browsed. The lowest level corresponds to the representation of the content only by using the associated visual codebook. Anyway, also this rough representation gives a quick idea of the color patterns that characterize the images of the given collection. It is also possible to visualize the content through the quantized version of the most representative image in the collection (second line in Fig. 8) which is the image that has been quantized by using the highest number of different visual codewords. Data can also be browsed in more detailed views

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

TABLE I
CLASSIFICATIONS RESULTS.

	Beach	**Mountain**	**Boat**	**House**
Beach	7	0	1	0
Mountain	0	7	0	1
Boat	0	1	7	0
House	0	0	0	8

according to the available quality layers.

C. Retrieval and Classification Capabilities

Beyond coding and fast browsing, the ability of automatically assigning a new image to a collection in the database has been also evaluated. In principle, this can be used to potentially avoid the proliferation of new collections but also to find collections visually similar to a query image. The image classification task has been realized by using the similarity estimation, described in Section II-B, which relies on the cross comparison of visual codebook. The considered test set is composed by four classes of images similar to those presented in Section III-B. Analyzing the confusion matrix reported in Tab. I the classification results are quite good, even if at the moment they are only indicative and a wider training and testing is in progress.

IV. CONCLUSION

This paper proposes an efficient method for scalable coding of collections of similar images, with embedded low-level visual descriptors. The key aspect that characterizes the proposed approach is the ability to embed in the compressed stream visual descriptors that possibly allow for a new scalability dimension. This paper also shows, how it is possible to enable browsing of image collections at different levels of detail: visual codebook, thumbnail, and images, by extracting and decoding the corresponding portion of the original bit stream. It is also possible to get a rough classification of the similarity of new images with respect to the collections in the repository. The coding efficiency is also good, as demonstrated in the experimental evaluations. Despite the promised potentials, several issues should be investigated in order to generalize and further enhance the presented results. The most difficult aspects will concern the generalization of the proposed approach to other application, and the determination of new low level descriptors useful for both data compression and data retrieval. In this respect, an analysis of the performance degradation produced by increasing the number of collections in the database should also be performed.

REFERENCES

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

[2] R. W. Picard, "Content access for image/video coding: "the fourth criterion"„" MIT Media Laboratory Perceptual Computing Section, Cambridge, USA, Tech. Rep. 295, 1994.

[3] A. Hanjalic, R. Lagendijk, and J. Biemond, "Efficient image codec with reduced content access work," *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, vol. 3, pp. 807–811 vol.3, 1999.

[4] G. Schaefer and G. Qiu, "Midstream content access of visual pattern coded imagery," *Computer Vision and Pattern Recognition Workshop*, pp. 144–144, 27-02 June 2004.

[5] G. Qiu, "Embedded colour image coding for content-based retrieval," *Journal of Visual Communication and Image Representation*, vol. 15, no. 4, pp. 507–521, 2004.

[6] C. Elkan, "Using the triangle inequality to accelerate k-means," in *ICML*, 2003, pp. 147–153.

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

[8] C. Saraceno and L. R., "Indexing audio-visual databases through a joint audio and video processing," *International Journal of Imaging Systems and Technology*, vol. 9, no. 5, pp. 320–331, 1998.