IMAGE CODING WITH FACE DESCRIPTORS EMBEDDING

Alberto Boschetti[1]*, Nicola Adami*[1]*, Riccardo Leonardi*[1]*, Masahiro Okuda*[2]

[1]Signal and Communication Lab, University of Brescia, Italy
[2]VIGpro Laboratory, The University of Kitakyushu, Japan

ABSTRACT

Content descriptors, useful for browsing and retrieval tasks, are generally extracted and treated as a separate entity with respect to the nature of the content itself. At the same time, conventional coding processes do not take into account information carried out by content descriptors. Content descriptors are closely related to the content itself, and they potentially can be used to exploit redundancy in entropy coding processes. Embedding content descriptors in the bitstream can reduce content description extraction load, and at the same time, it can reduce the rate associated to the compressed content and its description.

In this paper an effective implementation of this approach is presented, where image descriptors are actively used in the coding process for exploiting redundancy. First of all, image areas containing faces are detected and encoded using a scalable method, where the base layer is represented by the corresponding eigenface, and the enhancement layer is formed by the prediction error. The remaining areas are then encoded by using a traditional approach. Simulations show that achievable compression performances are comparable with those provided by conventional, making the proposed approach very convenient for source coding and content description.

Index Terms— eigenface, descriptors, scalable image coding

1. INTRODUCTION

Many services and applications (e.g. Facebook, Flickr and Picasa) offer the possibility to store, catalog and organize collections of pictures. According to [1] and [2] the uploading rate of pictures online is increasing month to month (currently Facebook receives over than 2.5 billion photos uploaded every month, in 2007 the rate was only 240 millions/month). One important feature, implemented in almost all these web services, is face recognition. It is used for tagging people appearing in a picture. This functionality, very appreciated by the users, makes it easier to select photos with some friends and share images among people in the same shot.

In the above application context, and every time browsing and retrieval operations are involved, images are stored by using formats that allow to add metadata, which contains the description of the scene to the compressed content itself (e.g. face locations and descriptors, tags, exif data and geolocalization). It is important to note that storing content descriptors, automatically extracted from the image as an independent entity, even if in the same data stream, in general increase redundancy. Contrarily, extracting content descriptors every time they are needed would require unwanted extra computation. Given these considerations, it is evident how, for example, it would be convenient if faces descriptors were also used to exploit redundancy in the image compression process. With respect to conventional approach, this strategy has several advantages:

- descriptors are fundamental elements of the compressed bitstream, which provides consistency between descriptors and the content they are referring to;
- considering the content and its descriptors, there is a global redundancy reduction;
- it allows a fast access to the descriptors with only partial decoding.

To evaluate the feasibility and the benefits provided by such a system, an original coding system for images which contains faces is hereafter proposed. According to this method, faces detected in a given image are at first roughly encoded by adopting PCA/eigenfaces based technique, while the residual image (reconstruction error) is compressed with JPEG2000 classical technique. By using this architecture, faces (the image descriptors) can be retrieved using only a dot product operation; hence, a fast access to the content description is guaranteed.

2. RELATED WORK

The main idea of this contribution has been derived form the concepts introduced by Picard [3], summarized in Figure 2. In this new paradigm of image coding, the user may be able to access and modify the semantic elements of a compressed content without decoding the entire image, by accessing the so called *Midstream*. This operation can lead the user to easily query, retrieve and manipulate content elements. The *Midstream Access Work* has been indicated by Picard as "The Fourth Criterion" because it is an additional optimization criteria to the three considered in a conventional image encoder (bit rate, distortion and complexity).

Old scenario
picture → code → (STORE,SEND) → decode → picture

New scenario
picture → code → (STORE,SEND) → decode → picture
⇕
CONTENT ACCESS

Fig. 1. Picard's Fourth criterion

In [4] the author achieves the midstream access capability through multiple conventional image coding methods: segmentation-based image coding (SBIC), vector quantization (VQ), and coloured pattern appearance model (CPAM). In this contribution descriptors are two: the joint probability distributions of the segmented region sizes, and their achromatic and chromatic spatial patterns' codebook indexes.

In [5] the image description is based on colour and shape (edges). These descriptors can be quickly retrieved from the stream

because the images are coded using the Colour Visual Pattern Image Coding (CVPIC) compression method. The author shows that this architecture is efficient for image retrieval in large databases.

Descriptors have been used in [6] and [7] for obtaining respectively a scalable image encoder and a shot–based video encoder. In both solutions, the description of the image is composed by the visual codebook, obtained by vector quantization of image blocks. Each image or frame is then decomposed into a series of indexes and the corresponding prediction error. By using these visual descriptors, authors showed that it is possible to easily retrieve and browse image collection characterized by similar visual patterns. Moreover, the selected descriptor allows to decode and visualize the images in a quality scalable fashion.

3. THE PROPOSED CODING SCHEME

The overall encoding process is accomplished in two main steps: the first stage concerns all the operations needed to train a learning machine based on eigenfaces (explained in 3.1), while the real encoding process is realized in a second stage. The produced output is formed by: a series of coefficients and a residual image. The formers are used as descriptors for face recognition and to build a prediction signal of the face; the latter is needed to reconstruct, lossy or lossless, the original input image.

3.1. Eigenfaces technique

The eigenface technique is nowadays widely used for face recognition purposes. As stated in [8] it is a variation of the PCA (Principal Component Analysis) method, applied to the faces in the images. The goal of this process is to generate a reduced set of eigenvectors which can describe the principal components of the input faces (named eigenfaces).

The result of the Principal Component Analysis technique is given by the most relevant K eigenfaces ($[\varphi_1, \varphi_2, \ldots, \varphi_K]$) and the average face image.

3.2. Training

The training scheme of the proposed architecture is shown in Figure 2.

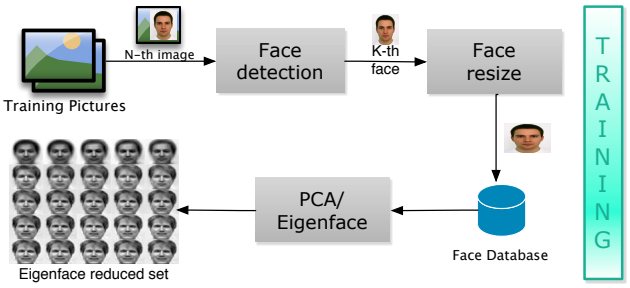

Fig. 2. Training chain

For each image in the training set, the faces contained in it are extracted with a face detection algorithm. The output of this initial step is a list of rectangles: every face in the image is described by its displacement and size within the picture. Successively, the faces are resizes to a standard size ($C \times R$ pixels) and converted to a grayscale image.

When all the faces of the training set have been processed and placed in the database, eigenfaces are generated using the process described in 3.1. Then, the most relevant K eigenfaces are selected as projection base, together with the averaged face used for input normalization. Each element of this projection base is a vector of $C \times R$ floating point values.

It is very important to train the system by using different faces. It is needed to have at least a few training samples for each sex, race, age, position, illumination and rotation of the possible face. In such way the K eigenvectors of the base contain more information, and they can be used for providing a better reconstruct approximation of different type of faces. It has to be noted that different strategies can be applied to eigenfaces extraction, according to the considered application. In this work we limit to the case where a unique projection base, available both at the encoder and at the decoder, is used for retrieval and coding of all faces detected in images. Another strategy could be, for example, to generate a projection base for any given collection of similar images and include the eigenfaces in the compressed bitstream. In this case the projection base, which is essential to reconstruct the original signal, could also be used to cluster image collection containing similar faces.

3.3. Encoder

The encoder and decoder scheme are shown in 3[1]. Initially faces are detected in the input image. Assuming that N faces are found; therefore, a list of N rectangles is provided by the face detection system. Each face is then resized (for matching the eigenface size) and converted to its grayscale representation. Successively, every face is projected, after the subtraction of the average–face, on the base of eigenfaces, obtaining a set of K representative coefficients. They are then quantized to the nearest integer and outputted together with the rectangle description. So far, for each face in the input image, a complete description is obtained; in fact, by using only these data, it is possible to create an approximation of the original face (in the exact position within the input image) and also to perform an automatic face recognition.

The next block reconstructs all the faces by using the above coefficients and reversing the operations implemented during face projection. Clearly, these reconstructed signals are good approximations of the originals, and so they are used as predictors. Predicted faces are then subtracted from the original ones, generating a residual image. All this operations are performed only on the luminance channel leaving chrominance unaltered.

The last step concerns entropy coding: the 3–channels image is compressed by using JPEG2000 while metadata are instead placed in a XML file, and then compressed using a lossless compression algorithm.

3.4. Decoder

The decoder side requests less computational power compared to the encoder part. The first step of the decoding chain is composed by the face reconstruction. In order to obtain the face predictors in the image, it is needed to decompress the metadata file, project the faces coefficients on the eigenfaces base and reshape the faces to fit their original size. It has to be noted that face coefficients and locations contained in metadata can also be used for fast content access and retrieval.

[1] Input image took at klabs.org/home_page/columbia

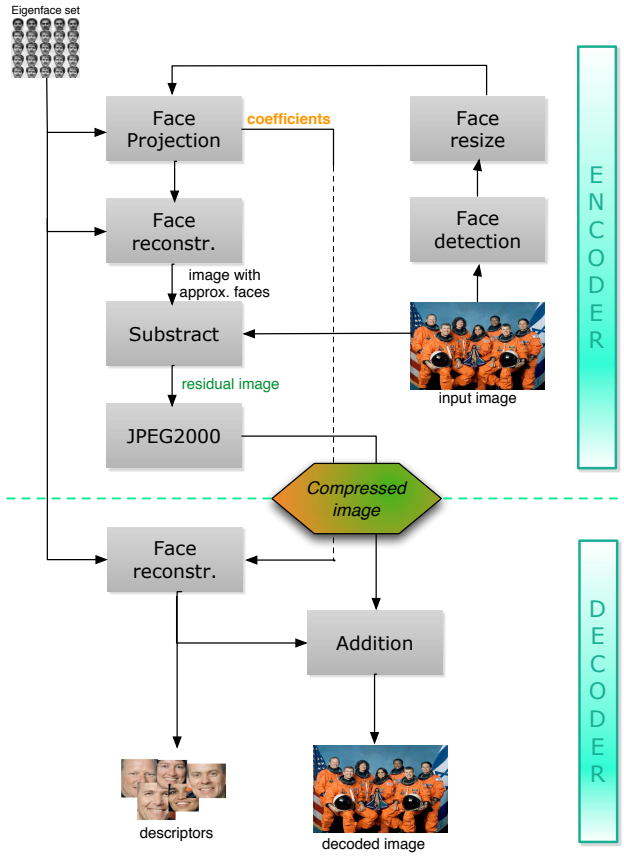

Fig. 3. Scheme of the encoder-decoder

In order to reconstruct the input image, the residual image is then decompressed. Successively, the predicted faces and residuals are recombined.

4. DISCUSSION

In this section, system parameters selection and performance evaluation are presented and discussed.

4.1. Training Data and Parameters selection

In order to correctly generate the projection base with a large range of details, different type of faces must be placed in the training set. For this purpose a combination of face–oriented (*Yale Database B* [9] and *AT&T Database of Faces* [10]) and general purpose images sets ([11] and [12]) have been used. The second kind of databases were chosen because they contain real–world images, not only related to faces. This guaranties that training is performed with a wide-spectrum of faces with different sex, race, age, position, illumination and rotation parameters.

The face detection algorithm, used in this work, has been derived from primitives provided by OpenCV (based on the algorithm [13]).

A key parameter for designing a good projection base is K, i.e. the number of the eigenvectors in the base. We chose it accordingly with the energy captured by the eigenvalues. In Figure 4 are shown both the eigenvalues energy and the cumulative sum of them (only the first ones are shown). We chose the parameter K so that the first K coefficients can catch at least the 90% of the face energy, so in our experiments it is resulted to be $K = 26$. By setting this value, the eigenface base size (K multiplied for the number of pixel in the face) and the number of coefficients stored for each face after projection are automatically determined.

For encoding and decoding of images and residuals, Kakadu, a JPEG2000 compliant encoder, has been used. It allows to produce quality scalable code-stream which is a very desirable property.

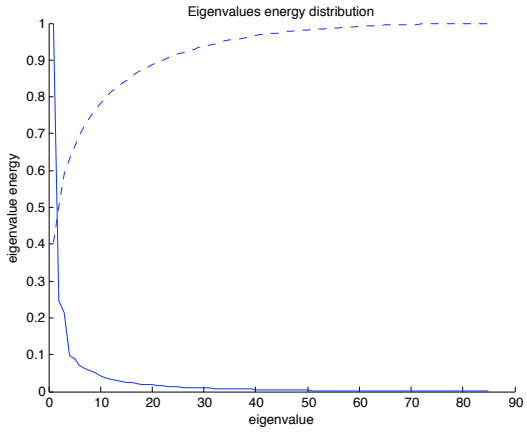

Fig. 4. Eigenvalues energy (line) and cumulative sum of them (dotted line)

4.2. Performance

In order to evaluate the coding performance of the proposed system, we compared the PSNR of the luminance of several image decompressed with the proposed method, with the results obtained by using conventional JPEG2000. The global rate, of our architecture, is computed as follow. Having some images to encode, the average rate per pixel of a generic picture with P pixel and F faces detected can be computed as the summation of three terms[2]:

$$mean(r)[bpp] = \frac{C[K \cdot FS]}{NP} + \frac{C[F \cdot (R+K)]}{P} + J2KR$$

where:

- K is the size of the eigenface vectors in the base;
- FS is the eigenface length (pixel);
- NP is the total number of pixel of the images to code;
- R is the byte count needed to store the rectangular description of faces (usually 4 integer values);
- $C[\cdot]$ is the lossless compression used to store the database and the metadata file;
- $J2KR$ is the rate of JPEG2000 compression encoder.

For each coded image, the global rate is the sum of the eigenspace contribution on the current picture (with more than 10 images to encode, this term becomes negligable), the metadata contribution (composed by both coefficients and rectangle descriptions) and the coded residual image.

[2]classical 8–bpp are considered

As shown in Figure 5 the performance of the proposed algorithm are similar to the pure JPEG2000 coding. Obviously, as previously stated, the proposed encoder allows to directly retrieve, in the compressed domain, the description of the image, without decoding the entire picture (as would be done using a traditional encoder).

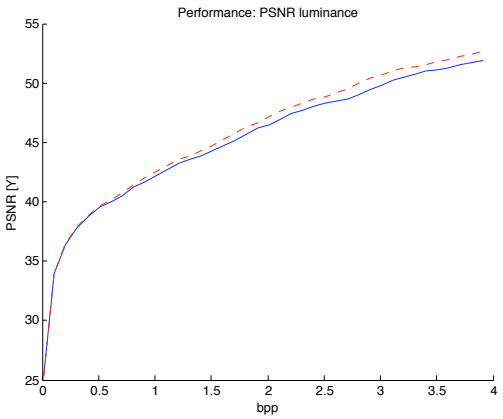

Fig. 5. Performance comparison (PSNR) between the proposed algorithm (line) and a pure JPEG2000 encoder (dotted line)

4.3. Example

An example of reconstructed image is shown in Figure 6. It is an enlargement on two faces when the compression factor is very high (0.11 bpp). Above the faces, their descriptors projected on the eigenface's space are shown.

Fig. 6. Two compressed faces (residual image coded at 0.11 bpp), and their fast accessible descriptors

5. CONCLUSION

A new coding approach, based on descriptors, has been introduced in this contribution. The main goal of this architecture is the possibility to access content descriptions in the compressed domain without extracting the entire content. To achieve it, descriptors are actively used in the encoding process.

Future works will concern the coding of the different parts of the picture by using optimal ad–hoc encoding methods (face with eigenfaces, bodies with another method, etc.). Moreover, retrieval evaluation will be performed by using the embedded descriptors.

6. REFERENCES

[1] Flickr Blog, "http://blog.flickr.net," .

[2] Facebook Blog, "http://blog.facebook.com," .

[3] Rosalind W. Picard, "Content access for image/video coding:the fourth criterion," Tech. Rep. 295, MIT Media Lab, 1994.

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

[5] Gerald Schaefer, "Midstream content access of visual pattern coded imagery," in *Conference on Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04.*, 2004, p. 144.

[6] Nicola Adami, Alberto Boschetti, Riccardo Leonardi, and Pierangelo Migliorati, "Embedded indexing in scalable video coding," *Multimedia Tools and Applications (MTAP), Special Issue on Content-Based Multimedia Indexing*, vol. 48, no. 1, pp. 105–121, 2010.

[7] Nicola Adami, Alberto Boschetti, Riccardo Leonardi, and Pierangelo Migliorati, "Scalable coding of image collections with embedded descriptors," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Chania, Crete Island, Greece, 8-10 October 2008.

[8] Matthew Turk and Alex Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, pp. 71–86, January 1991.

[9] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[10] F.S. Samaria and A.C. Harter, "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, Dec. 1994, pp. 138 –142.

[11] Jana Machajdik and Allan Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the international conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 83–92, ACM.

[12] Psychological Image Collection at Stirling (PICS), "http://pics.psych.stir.ac.uk," .

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001.*, 2001.