A PROPOSAL FOR HIGH COMPRESSION OF FACES IN VIDEO SEQUENCES USING ADAPTIVE EIGENSPACES [1]

Luis Torres   *Daniel Prado*

Technical University of Catalonia
Barcelona, Spain
{luis, aldanip}@gps.tsc.upc.es

ABSTRACT

This paper presents a proposal for a novel video coding scheme intended to encode human faces in video sequences at very high compression using a recognition and reconstruction approach. The scheme is based in the well-known eigenspace concepts used in face recognition systems, which have been modified to cope with the video compression application. Preliminary draft results are presented which look very promising. The authors want to emphasize that this paper is intended as a discussion paper as our work is at a preliminary development stage and no final results can yet be provided.

1. INTRODUCTION

Image and video coding are one of the most important topics in image processing and digital communications. During the last thirty years we have witnessed a tremendous explosion in research and applications in the visual communications field. There is no doubt that the beginning of the new century revolves around the "information society." Technologically speaking, the information society will be driven by audio and visual applications that allow instant access to multimedia information. This technological success would not be possible without image and video compression. The advent of coding standards, adopted in the past years, has allowed people around the world to experience the *digital age*. However, and in spite of all this effort, there are some applications that still demand higher compression ratios than those provided by state of the art technologies.

In particular, and due to its high applicability, there is a need to provide novel compression schemes to encode faces present in video sequences. Although the new standards H.263+ [1] and the synthetic part of MPEG-4 [2] along with other model-based proposed schemes [3] achieve high compression ratios for this particular application, we still believe that further compression is needed, among others, for mobile and video streaming environments.

It is in this context that we present a novel scheme to encode faces in video sequences based on an eigenspace approach. The eigenface concept for still image coding has been already presented in a face recognition framework in [4]. However, to the best of our knowledge, our approach is original and adapts the eigenspace to the video sequence to take into account the different poses, expressions and lighting conditions of the faces.

Section 2 presents an introduction to the topic of very high compression and the basic eigenspace concepts on which our scheme is based. Section 3 is our main contribution and introduces the general scheme for video sequences taking into account adaptive eigenspaces.

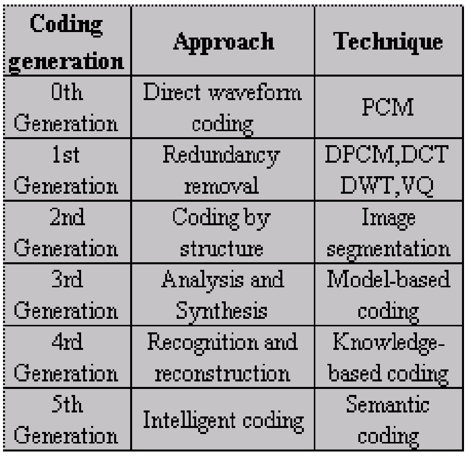
2. IMAGE CODING THROUGH RECOGNITION

2.1 Introduction

Many proposals have been made in the last years for image and video coding. A broad classification of all possible approaches can be seen in Figure 1 [5]. This classification is presented here in order to fully understand the characteristics of our proposal. Regarding the state of the art standards for high compression, H.263+ [1] provides a block-based redundancy removal scheme for doing low to high data rate robust compression. In addition, MPEG-4 [2] combines frame-based and segmentation-based approaches along with model-based video coding in the facial animation part of the standard which allows efficient coding as well as content access and manipulation [6]. It can be said that H.263+ and MPEG-4 represent the state of the art in video coding [7].

[1] This work has been supported by Grant TIC98-0422 of the Spanish Government.

Our proposal relies on fourth generation video coding techniques based on recognition and reconstruction [5]. Recognition and reconstruction approaches rely on the understanding of the content. In particular, if we know that an image contains a face, a house, and a car, recognition techniques to identify the content can be developed as a previous step to coding. Once the content is recognized, content-based coding techniques can be applied to encode each specific object. MPEG-4 provides a partial answer to this approach by using specific techniques to encode faces and to animate them.



Coding generation	Approach	Technique
0th Generation	Direct waveform coding	PCM
1st Generation	Redundancy removal	DPCM,DCT DWT,VQ
2nd Generation	Coding by structure	Image segmentation
3rd Generation	Analysis and Synthesis	Model-based coding
4rd Generation	Recognition and reconstruction	Knowledge-based coding
5th Generation	Intelligent coding	Semantic coding

Figure 1. Image and video coding classification

2.2 Face coding through recognition and reconstruction

Let us simplify the visual content by assuming that we are interested in the coding of faces in a videoconference session. Let us assume that automatic tools to detect a face in a video sequence are available. Then, some experiments show that a face can be well represented by very few coefficients found through the projection of the face on an eigenspace previously defined [4]. The image face can be well reconstructed (decoded), up to a certain quality, by coding only very few coefficients.

Face coding can be done using a fixed eigenspace, what decreases the quality of the reconstructed image, or by adaptively changing the eigenspace according to the changes experienced by the face to be encoded as appears in the video sequence. Section 3 gives the details of the adaptive technique. Prior to coding, the corresponding face is detected and separated from the background using some segmentation technique.

Once the face has been detected and coded, the background remains to be coded. This can be done in many different ways. The simplest case is when the background is roughly coded using conventional schemes such as JPEG for static backgrounds (first generation coding) or MPEG-4 for moving backgrounds (second generation coding). If the background is not important, then it can even be not transmitted and the decoder adds some previously stored background to the transmitted image face.

2.3 Video coding using non-adaptive eigenspaces

Our coding technique is based on a face recognition approach, which has been modified to cope with the coding application [9]. It assumes that a set of training images for *each* person contained in the video sequence is previously known. Once these training images have been found (usually coming from an image database or from a video sequence), a Principal Component Analysis (PCA) is performed for each individual using the corresponding training set of each person. This means that we obtain a PCA decomposition for every face image to be coded. The PCA is done previously to the encoding process.

After the PCA, the face to be coded is projected and reconstructed using each set of different eigenvectors (called eigenfaces) obtained in the PCA stage. If the reconstruction error using a specific set of eigenfaces is less than a threshold, then the face is said to match the training image which generated this set of eigenfaces. In this case we code the recognized face by quantizing only the most important coefficients used in the reconstruction. The error image can be also coded and transmitted to the decoder to improve the quality or to have a scalable scheme. The size of the coded image has to be previously normalized for PCA purposes and then denormalized at the decoder.

It is clear that the corresponding eigenspace of each person has to be transmitted previously to the decoder. This can be done by transmiting the training images and reconstructing the eigenspace at the decoder side. However, as the number of training images is expected to be reduced, this can be done using conventional still image coding techniques such as JPEG and no significant increment in bit rate may be expected.

Figure 2 shows the approach to face coding using a non-adaptive eigenspace. Let us emphasize that this scheme always uses the same training images and no adaptation to the video sequence content is made. Section 3 will present a general eigenspace approach for video coding that adapts itself to the important changes of expression or new individuals likely to appear through the sequence.

Some preliminary results of the coding scheme using the non-adaptive eigenspace approach are presented now for completeness purposes.

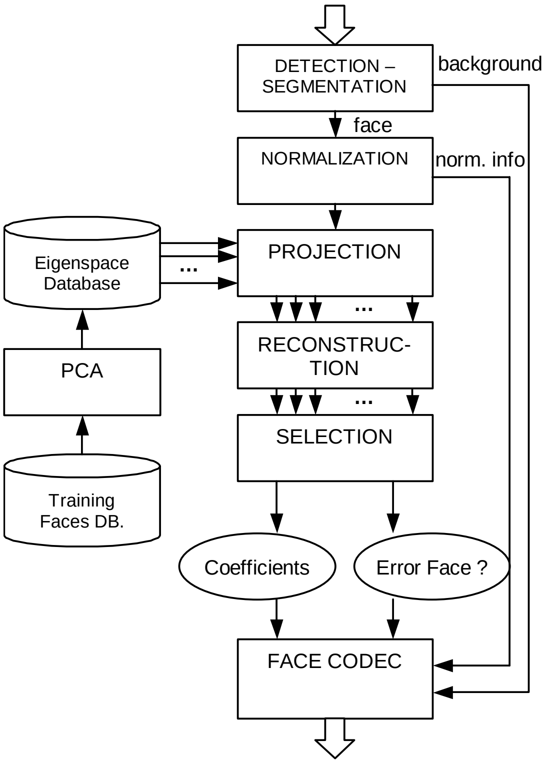


Figure 2. Block diagram of the non-adaptive eigenspaces approach for video coding

Figure 3 shows eight consecutive original frames of the *news11* sequence defined in the MPEG-7 content set. Each image is 58 x 68 pixels. The images have been automatically segmented using the face detection scheme proposed in [8]. Figure 4 shows the 5 images used to generate the eigenspace. The number of images used to generate the eigenspace can be changed, but 5 seems to provide a good compromise.

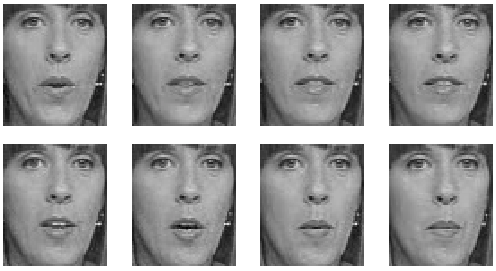


Figure 3. Original images of the sequence news12 from left to right and top to bottom.

Figure 5 shows the encoded images. The first image of the sequence has been coded using JPEG. As the face changes along the time, the eigenspace should be updated. In this example the last image can not be coded using the projection coefficients and JPEG has been used to encode this image. The intermediate images have been coded using only 5 coefficients which can be coded using 2 bytes per coefficient. This gives 0.022 bits/pixel for each image. No effort has been done to encode the error image. Some errors can be noticed as the mouth expression is not always well represented.

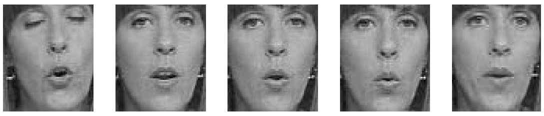


Figure 4. Original images used to generate the eigenspace

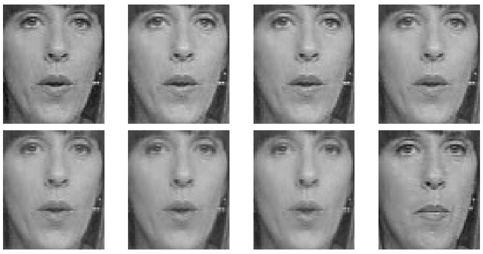


Figure 5. Coded images

3. VIDEO CODING USING ADAPTIVE EIGENSPACES

3.1 Introduction

The non-adaptive eigenspace approach to face coding has shown that if the face image does not experience major changes, a single eigenspace can be used. However, if the image to be encoded is not *very* similar to any of the corresponding training images, then the decoded image decreases in quality because is only similar to a combination of the training images. In this section we propose an eigen scheme that adapts itself to the face content appearing in the video sequence. Figure 6 shows our proposal. Notice that a related technique has been presented in [10] although the approach is significantly different.

As in the non-adaptive scheme, faces obtained from the segmentation process are projected and reconstructed using every eigenspace contained in the eigenspace database. The reconstruction with the minimum error is chosen. Ideally, the face is coded with only a few coefficients resulting from this process. Nevertheless, we can predict that any important change in the expression of the face that is not included in the training images, would lead the system to poor quality reconstruction.

In order to overcome this problem, we have designed a fall back mode system, which consists of a quality of reconstruction evaluation block followed by an upgrade mechanism of the coder and decoder eigenspace databases. If the error obtained from the reconstruction process is low, only the coefficients of the reconstruction are sent to the decoder.

If an important change in the expression of the face leads to a high error result, but still enough to allow recognition, that face will be coded using one of the available coding techniques like JPEG. In addition, the same face is added to the training images of the recognized person, and the corresponding eigenspace is so recalculated. In this way the faces of that person contained in the following frames, will be better represented using the upgraded eigenspace. This provides a basic approach for updating the eigenspace.

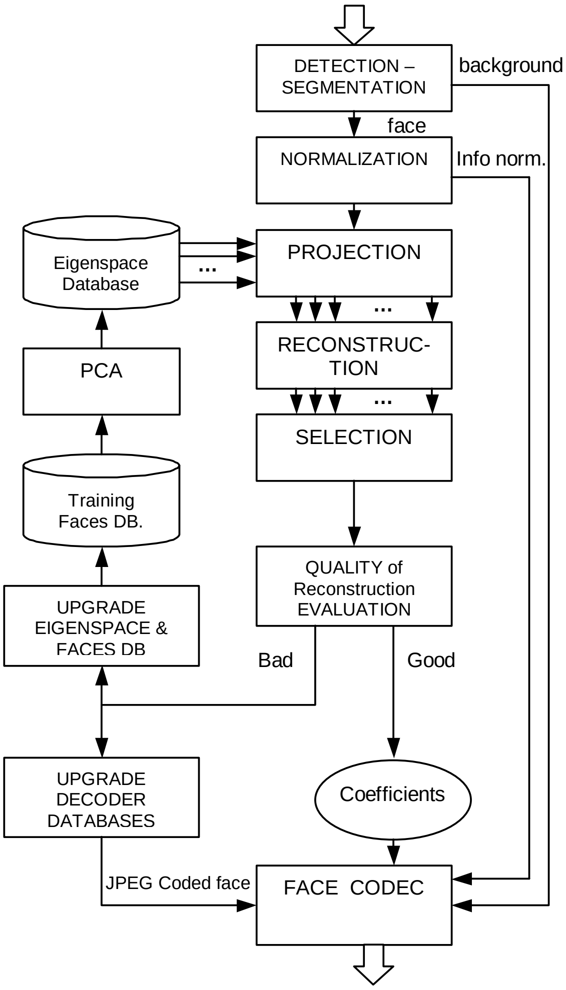


Figure 6. Block diagram of the adaptive-eigenspaces approach video coder

It also can occur that the reconstruction error is so high that the person among those contained in the training database cannot be recognized. In this case, a new eigenspace is dinamically created by obtaining a set of new images from the sequence and sending them to the decoder using any well-known compression scheme. Once this is done, the following faces can be potentially coded using the eigen technique.

3.2 Eigenspace database updating system

The updating PCA system we propose works as follows: let us assume the eigenfaces of a person have been obtained by calculating the PCA of a single-person using N training images. The encoder starts coding the faces of the video sequence by transmitting the coefficients of the projection of the original faces over the eigenspace. In the coded sequence, the transmitted frames using only the coefficients of projection are called ‘P’ (for projection) frames. The frames used to update an eigenspace are called ‘U’ (for ‘update’) frames. The system continuosly evaluates the mean square error of the reconstruction. When the error increases over a specified limit, the eigenspace has to be updated at both the coder and the decoder.

A first updating system approach can be designed by substituting the *oldest* face in the group by the face which is being now coded. However, this system has an obvious problem. We are wasting the face that is updated in the substitution process. It would be more interesting to store the information of any existing frames, as the images of a ‘talking-head’ sequence present short-term and long-term correlation. To illustrate this, let us assume we have a 10 face training database, formed mainly with the first frames of a sequence where the person has a *happy* expression. During the encoding process, if the person changes its expression to a *sad* one, the system will update the face database. But at the same time, if the oldest face is substituted, part of the database will be destroyed and the system will not be able to reconstruct anymore *happy* faces using only projection coefficients.

There are two solutions to this problem. The size of the training images set can be increased without eliminating any face each time an ‘U’ frame has to be sent (unless a maximum number of training images is fixed). However, as the system increases the size of the training images, the eigenspace generation becomes more and more time-consuming, and the system will not be able to perform the necessary operations in a reasonable time. The other solution consists of a multiple low-dimension eigenspace database. The system starts coding using a single eigenspace as before. When an update-frame is transmitted, the training group is duplicated, and the substitution process is performed over one of the resulting groups. A new PCA is then calculated over the new sub-group and the system obtains two eigenspaces. The following frames are reconstructed with the two (or more) existing eigenspaces and the one with the minimum reconstruction error is selected. A maximum number of eigenspaces per person is fixed, and when it is reached, the updates are made by direct substitution of the oldest face of the selected eigenspace. This system has an advantage compared with the single PCA described before. It is far less expensive, computationally speaking, to manage 10 eigenspaces of dimension 5 than a single eigenspace of dimension 50. Firstly, because every time an update must be done, the first system has to calculate only an eigenspace of dimension 5, whereas the dimension is 50 in the second one. Secondly, because the PCA is a much more computationally intensive process than the projection and reconstruction one. So to calculate 10 reconstructions per frame is not as time consuming as calculating the eigenspace.

4. PRELIMINARY RESULTS

Our scheme is at an early stage of development. Nevertheless, preliminary results of the *news11* coded sequence of 150 frames will be presented. The results are shown in Table 1.

Original format	58 x 68, Y component, RAW (3944 bytes/frame)
% of U frames	7.8 %
% of P frames	92.2 %
Average PSNR of 'P' frames	23.4 dB
Dimension of the PCA	19
# of coefficients transmitted/'P'frame	5
Size of U frames	1000 bytes aprox.
Mean bitrate	98 bytes/frame (19.6 Kbps at 25 fr/s)
Compression factor	40

Table 1. Results of encoding 150 frames of the sequence *news11*

It has to be noticed that the bit stream may present strong fluctuations if the eigenspaces have to be continously adapted, thus decreasing the efficiency of the proposed scheme.

5. ERROR IMAGE CODING

In order to improve the quality of the reconstructed image, at the expenses of increasing the overall bit-rate, the error image resulting from the reconstruction can be coded and transmitted. This alternative can be useful both in the static eigenspace database system presented in Section 2.3 and in the adaptive-eigenspace system presented in Section 3.

To maintain a reasonable updating rate of e.g. 1 / 20 frames, we propose the use of an efficient static image coder that fits well with the nature of reconstruction-error images. Although we cannot present definitive results yet, we are working in a promising technique consisting in the use of quantization and JBIG coding [11] [12]. The main objective is to obtain a reasonable improvement in quality of the decoded images while maintaining the bi-trate increase as low as possible.

The size of the error images to be coded is the same as that of the reconstructed face, and has a pixel dynamic margin of [-255, 255] (9 bits/pixel). First, a low-pass filtering and downsampling process is performed resulting an image a quarter the size of the original. Second, the image is non-uniformly quantized into a 15-level image (1 level for '0' error value, 7 levels for positive values and 7 levels for negative values), which results in a 4 bits/pixel image. The quantization and dequantization process is adapted to the histogram of the input image. The compression factor from the downsampling and quantization process is 9. The image is then splitted into 4 bit-planes that are coded using a loss-less JBIG compressor [11] [12]. The coded error image resulting from the complete process can reach a minimum compression factor of 20. Although this could seem not much compared with other systems like JPEG, we have to consider that the size of the faces involved is quite small (tipically 60 x 70 pixels). This causes traditional coders like JPEG and also JBIG to offer less compression results, as the headers and tables of the coded file represent an important cost in the final bitstreams.

As JBIG is a coder based on the source statistics, the compression factors are better when the input error image is bigger. We have taken advantage of this by grouping the small error images into a bigger error-image-matrix of 2x2 or 3x3 images. This can improve the compression by a factor of 2.

Figure 7 presents some results of the proposed error coding system. First image on the left is an original frame of the MPEG-7 content set sequence *news11*. Next one is the reconstructed image using the projection over the eigenspace system. Third and fourth images are the original and coded error image respectively. The last image is the final decoded image using the transmitted error. The improvement in visual quality is noticeable. The size of the coded error image is 420 bytes. Coded as a part of a 3x3 error-image-matrix the size would be about 250 bytes.

In order to compare JBIG and JPEG coders for error prediction, Figure 8 presents some results using conventional JPEG. The size of the coded error image (fourth image) is 744 bytes, and the quality is appreciably worse due to strong block-effect.

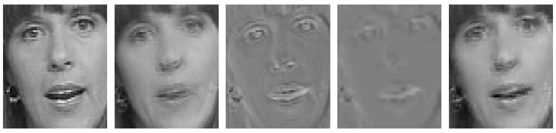


Figure 7. Error image coding using quantization and JBIG.

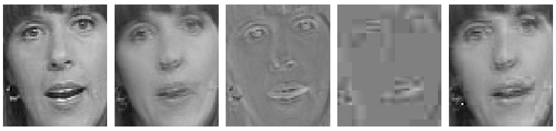


Figure 8. Error image coding using JPEG.

Table 2 presents the results of coding the *news11* sequence, but with error image transmission. Error images are only transmitted when the reconstructed image PSNR is less than 21 dB. This means that, for this example, 25.2 % of reconstruction error images are coded and transmitted. Notice the increase in mean PSNR compared with Table 1 and the better subjective impression of the decoded image (Figure 7). Frame downsampling can be done to decrease the bit-rate. No attempt has been made to encode the background image associated to the face image.

Original format	58 x 68, Y component, RAW
% of U frames	7.8 %
% of P frames	92.2 %
Average PSNR of 'P' frames	24.1 dB
Dimension of the PCA	19
# of coefficients transmitted/'P'frame	5
Mean bitrate	157 bytes/frame (31.4 Kbps at 25 fr/s)
Compression factor	25

Table 2. Results of encoding 150 frames of the sequence *news11*, adding error coding.

6. CONCLUSIONS

An adaptive eigenspace approach for encoding moving faces has been presented. Very preliminary results look promising. Although the obtained bit-rate is not yet competitive, the eigenspace approach looks promising and more work needs to be done in order to check the validity of the proposed technique.

7. REFERENCES

1. G. Côté, B. Erol, M. Gallant, and F. Kossentini, "H.263+: Video coding at low bit rates", *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 8, no. 7, November 1998.
2. ISO/IEC ISO/IEC 14496-2: 1999: Information technology – Coding of audio visual objects – Part 2: Visual, December 1999.
3. P. Eisert, B. Girod, "Analyzing facial expressions for virtual videoconferencing", *IEEE Computer Graphics and Applications,* Vol. 18, No. 5, pp. 70 - 78, 1998.
4. B. Moghaddam, A. Pentland, "Probabilistic visual learning for object representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 696-710, July 1997.
5. H. Harashima, K.Aizawa, and T. Saito, "Model-based analysis synthesis coding of video-telephone images – conception and basic study of intelligent image coding", *Transactions IEICE*, vol. E72, no. 5, pp. 452-458, 1989.
6. F. Pereira, "Visual data representation: recent achievements and future developments" *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Tampere, Finland, September 5-8, 2000.
7. L. Torres, E. Delp, "New trends in image and video compression", *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Tampere, Finland, September 5-8, 2000.
8. F. Marqués, V. Vilaplana, and A. Buxes, "Human face segmentation and tracking using connected operators and partition projection", *Proceedings of the IEEE International Conference on Image Processing*, Kobe, Japan, October 1999.
9. L. Torres, L. Lorente and J. Vilà, "Face recognition using self-eigenfaces", *International Symposium on Image/Video Communications Over Fixed and Mobile Networks*, Rabat, Morocco, pp. 44-47, April 2000.
10. W E. Vieux, K. Schwerdt and J.L. Crowley, "Face-tracking and Coding for Video-Compression", First International Conference on Computer Vision Systems, January , 1999.
11. International Standard ISO/IEC 11544:1993 and ITU-T Recommendation T.82 (1993), "Information technology - Coded representation of picture and audio information - progressive bi-level image compression"
12. Markus G. Kuhn, JBIG-KIT implementation. <http://www.cl.cam.ac.uk/~mgk25/download/jbigkit-1.2.tar.gz>