Principal Component Coding of Mouth Sequences

*D. Shah, S. Marshall, W.J.Welsh**

Signal Processing Dvision, Dept. of EEE, University of Strathclyde, Glasgow G1 1XW, U.K.
* British Telecom Research Labs, Martlesham Heath, Ipswich IP5 7RE, U.K.

ABSTRACT

Facial Features of a videophone image sequence are coded using the traditional technique of Principal Components (PCs). One of the main advantages of this method is that it makes use of a high apriori knowledge of the images as well as exploits the correlation between pixel values. The sequences are coded through a hybrid wire frame model in which the head and shoulder images are coded using a low number of bits. A box is placed around the mouth and eye regions which are coded at a higher bit rate using PCs. Since the image sequence is similar in its overall configuration, this scheme gives a very high compression ratio of around 500:1.

1. INTRODUCTION

The present day teleconferencing systems store and transmit images in the digital medium. These digital representations generate huge amounts of data. Coding and storing in a straight forward way would require high capacity digital transmission channels and large amounts of memory. As cost is proportional to capacity, it is desirable to keep the quantity of data as low as possible.

This can be achieved through data compression. Significant compression is obtained by exploiting the correlation between pixels of an image. The image compression techniques such as DCPM, transform coding, vector quantization, etc. are considered as conventional techniques. These above mentioned techniques achieve some compression but when compared to the requirements of Study Group XV of CCITT [1] for low bit rate transmission at multiples of 64 kbits/s (known as model H. 261), the algorithms are stretched to their limit. Hence, the algorithms produce low quality images when there is significant motion due to low frame rate. Using the technique of Model Based Coding, there is the opportunity of exploiting much more of the redundancy in a sequence. It has been noted by Welsh [9] that using the same rate of transmission, image quality of the sequence is improved. One of the difficulties lies in coding the small scale changes that occur frequently in a head and shoulder sequence. A specific problem is the loss of lip synchronization.The method presented here makes use of a hybrid wire frame model [8] and the critical aspects of the picture, such as lips synchronization, can be coded efficiently by returning to the more traditional method of principal components. The advantage in using this method is that it makes use of a very high apriori knowledge of the images and hence it gives a huge compression. Another means of dealing with the problem of lip synchronization, would be to use speech signals to match the movements of the lips [10]. An approach which might be adopted would be to examine the speech signal in order to determine the periods when speech is present and when there is no speech. If it is assumed that the mouth will generally be closed when there is no speech and open otherwise, two different sets of PCs can be derived for these cases. In [8], a method was developed in which a single set of PCs were derived for all the mouth images in a training set. The motivation for the current work was the belief that using two sets of PCs would improve the quality of reproduced images. Hence these methods can be introduced as an enhancement to the current H.261 based codecs.

2. MODEL BASED CODING TECHNIQUE

The idea of Model based coding or Knowledge based coding was first introduced by Parke [7] in 1982. He proposed that a parameterized model can produce realistic images of human faces which could then be used for videophony. Later Yau[3] worked on a system for modelling a person's head. Welsh et al [9] used a three dimensional model for producing synthetic facial images.

The idea behind this technique is to produce models of scene objects and use them to form images. In other words, information is transmitted through parameters of the model. This is turn gives a better compression. In the case of videophone images, modelling of the head and shoulder images is required. Throughout a conversation, the subject's facial features change frequently but their overall appearance does not, except when the subject moves away and someone else takes their place. Hence

0-7803-0946-4/93 $3.00 © 1993 IEEE

considering the latter situation and with no introduction of new objects, most of the information can be transmitted in the first few frames with only motion information transmitted in the remaining frames. But this scheme produces artifacts when mouth and eyes are coded. So we propose a hybrid scheme where the head and shoulder images are transmitted at very low bit rates along with a box with eyes and mouth regions coded with more bits using the Principal Component method. Methods are being developed to track facial features such as eyes and mouth in a robust way.

3. PRINCIPAL COMPONENT APPROACH

The use of Principal Components for synthesizing and recognizing face images was suggested by Kirby and Sirovich [4] and later used elsewhere [5][6]. This idea has been used as the basis for image compression, by forming a set of eigenvectors corresponding to a set of sub-images of the mouth area. The set of mouth images are extracted from a sequence of face images of the same subject and are referred to as a training set. For the purpose of evaluating the scheme, the mouths are located manually. These images are then geometrically normalised so they have a consistent orientation and hence reduce the variation in the data as far as possible.

Using the normalised mouth images $I_1, \ldots, I_M$ where M is the number in the training subset, a set of difference images are obtained by subtracting the average mouth image from each image in the normalised set. The average image is calculated by

$$\bar{I} = \frac{1}{M} \sum_{n=1}^{M} I_n \qquad (1)$$

and each difference image is obtained by

$$\phi_i = I_i - \bar{I} \qquad (2)$$

This set of very large vectors is subject to PC analysis [2]. Using this difference set, an M x M matrix L is constructed and the M eigenvectors v_l of L are obtained, where

$$L_{mn} = \phi_m^T \phi_n \qquad (3)$$

The M orthogonal eigenpictures E_l are then calculated using the M eigenvectors v_l of L and the difference set ϕ_n:

$$E_l = \sum_{n=1}^{M} v_{ln} \phi_n \, , \; l = 1, \ldots, M \qquad (4)$$

This set of M eigenpictures span a basis set, which best describe the mouth distribution and the first few eigenpictures carry the maximum variation.

Having obtained the eigenpictures from the training ensemble, mouth images outside this set can be represented as follows:

The new image I' is transformed into its eigenpicture components by the operation

$$\omega_k = E_k^T (I' - \bar{I}), \quad k = 1, \ldots, M \qquad (5)$$

The pattern vector $\Omega^T = [\omega_1, \omega_2, \ldots, \omega_M]$ describes the contribution of each eigenpicture in representing the new mouth image. The new image can then be resynthesized using the representative set of coefficient and the eigenpictures. This is given by:

$$I' = \sum_{k=1}^{M} \omega_k E_k \, . \qquad (6)$$

4. RESULTS

Experiments were carried out for a set of 75 mouth images from a sequence of 200 head-and-shoulder images of the same subject. The mouth corners of each frame were manually located and a window of 96 x 80 pixel resolution was extracted. The first 40 frames were used to produce the eigenpictures and the following 35 frames were coded and reconstructed in various ways. The fidelity criterion used for measuring the image quality was the peak S/N ratio, which being a coarse measurement was not fully able to reflect the subjective quality of an image sequence.

In the first experiment 35 frames were coded and reconstructed using 5, 10 and 20 eigenpictures. It was found that the sequence resulting from 20 eigenpictures showed no significant improvement and that only about 1dB improvement in SNR was obtained when the number of eigenpictures was doubled. Table 1 shows the performance of this experiment. The compression ratio for the coded sequence was found to be 192. The compression ratio is found by dividing the original number of bits required for each image (i.e. 96 x 80 x 8) with the number of bits used to code the image (i.e. the coefficient precision multiplied by the number of eigenpictures).

A second experiment was carried out in which 10 eigenvectors were used and the pattern vector elements were quantized into 12 or 14 bits. This gave a huge compression ratio of 512 or 439 respectively without significantly reducing the image quality. No further improvement was obtained from using more than 12 bits per coefficient, which can be seen from Figure 1.

In the third experiment, mouth images were classified into open and closed mouth shapes. This was believed to give an insight into the benefits of incorporating speech signals into the processing. The determination of whether the mouth was open or closed is based on thresholding the distance between the upper and lower lips. The PC's were then obtained for the two sets and used to code new mouth images depending on whether it was open or closed. It was found that the sequences resulting from 8 and 10 eigenpicture had a blurred appearance. The sequence resulting from 12 eigenpictures was found to be of better quality which was reflected in a 1 dB improvement in the S/N ratio. The compression ratio for the coded sequence was found to be 160 and can be seen from Table 2.

In the fourth experiment the pattern vector elements were quantized into 8, 10 and 12 bits with the number of eigenpictures fixed at 12. It was found that though the SNR values did not differ significantly as seen from the graph, the sequence resulting from 8 or 10 bits per coefficient gave adequate image quality. This gave a huge compression factor of 640 and 512 respectively. Figure 2-4 show examples of the coded images in comparison with the originals.

A further experiment consisted of coding the eyes using the PCA technique. The experiments carried out were equivalent except that a window of 30 x 50 resolution was extracted from the face image. It was found that the eye images could be coded using fewer number of eigenpictures in comparison to mouth images. The reason was that nearly all the variation was carried in the first few eigenpictures. 35 new frames of the sequence were coded using different numbers of eigenpictures. Sequences obtained using six eigenpictures were shown to be of adequate quality. Using 8 or more eigenpictures the image sequences showed no improvement. Also from Table 3, it can be seen that there was no significant improvement in the SNR value when the number of eigenpictures were increased. The compression ratio without quantizing the coefficients, for the eye image of resolution 30 x 50 was 63.

Another experiment was carried out in which the number of eigenpictures was fixed at 6 and the pattern vector elements were quantized to various numbers of bits. Figure 1 shows the result of quantizing the coefficients. Using 3 or 4 bits per coefficient, a good quality image was obtained with compression ratios of 500 and 667 respectively. From the SNR values shown in Figure 1, it was observed that using more bits per coefficient resulted in no improvement in image quality. Fig 5 shows an example of an open eye in its original state and coded state.

5. CONCLUSION

An informal subjective assessment concluded that no perceptual improvement was made in the first two experiments, when using more than 10 eigenvectors and 12 bits per coefficient. In the second set of experiments, the mouths were grouped into open and closed sets and coded using two codebooks. It was observed that no gain in image quality was achieved, when more than 12 eigenpictures and 10 bits per coefficient were used. Using 8 bits per coefficient also gave an adequate quality image. Although the S/N ratio figures seem poor for this experiment, the subjective quality is not greatly affected as the S/N values suggest. Hence it is concluded that using different codebooks for different shapes of mouth, gave the same quality image sequence but a higher compression ratio when compared to using a single codebook. In the fourth example, the reproduction of the coded mouth is poor as that particular shape is not present in the training set. It is obvious that the quality of the reconstructed sequence depends on whether the training set is adequate or not. In the last set of experiments, depending on the image quality required, sequences generated using 3 or 4 bits per coefficients were of adequate quality and no improvement was obtained using more than 6 eigenpictures and 3 or 4 bits per coefficient.

REFERENCES

[1] CCITT: 'Reference Model'. CCITT SG XV (Specialist Group on Coding for Visual Telephony) COST 211 bis Paris, 1989.

[2] I. T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York Inc, 1986.

[3] J. F. S. Yau, N.D. Duffy, "A texture mapping approach to 3D facial image synthesis", 6th Annual Eurographics (UK) Conf. Sussex (6-8 April 88).

[4] L. Sirovich and M. Kirby, " Low-dimensional procedure for the characterization of human faces", J. Opt. Amer. A, vol 4 no. 3, 1987.

[5] M. Turk and A. Pentland, "Representing faces for recognition", MIT Media Lab Vision and Modeling Group tech. rep. 132, 1990.

[6] M. A. Shackleton and W. J. Welsh, "Classification of facial features for recognition", Proc. IEEE CVPR, Maui, Hawaii, June 3-6, 1991.

[7] F. I. Parke, "Parametrised models for facial animation", IEEE CG and appl. vol 12, nov 1982.

[8] W. J. Welsh, D. Shah, "Facial feature image coding using principal components", Electronics Let, vol 28, no. 22.

9] W. J. Welsh, S. Searby, J. B. Waite, "Model-based image coding", Br Telecom Technol J, vol8 no 3, July1990.

[10] W. J. Welsh, A. D. Simons, R. A. Hutchison, S. Searby, "Synthetic face generation for enhancing a user inter-

face", Proc. Image' Con 90, 1st Int Conf. on new image chains. 177-182, Bordeaux, France, Nov 1990.

ACKNOWLEDGEMENT

This work was initiated at BT Research Labs, Martlesham Heath, England during a study visit by Miss D. Shah, University of Strathclyde. BT also provided a bursary for travelling to ICASSP-93.

Table 1: Performance of PCA method for the mouth images without quantisation

No of PC's used	Aver of peak SNR over every frame	Compressio n ratio
5	27.911	384
10	**29.298**	**192**
20	30.241	96

Table 2: Performance of PCA method for the mouth images using two codebooks and without quantisation

No of PC's used	Aver. of peak SNR over every frame	Compression ratio
8	26	240
10	26	192
12	**27**	**160**

Table 3: Performance of the PCA method in the application of eyes (without quantisation)

No of PC's used	Aver. of peak SNR over every frame	Compression ratio
4	34.001	93.75
5	34.555	76.6
6	**34.998**	**62.5**
8	35.723	46.87
10	36.177	37.5

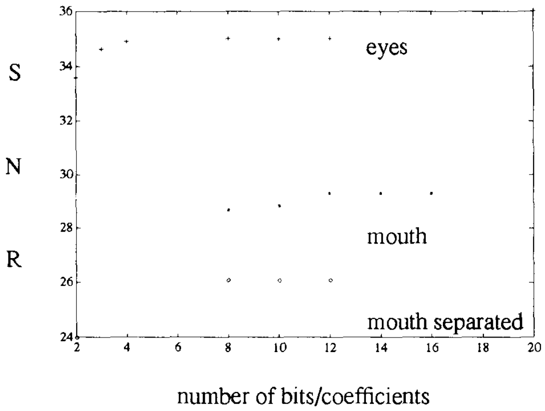

Figure1: Performance of the three sets of experiments in terms of no. of bits per coefficients and the average SNR ratios

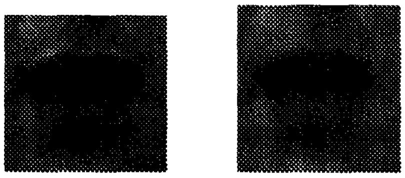

Figure 2. Original and coded mouth images (open)

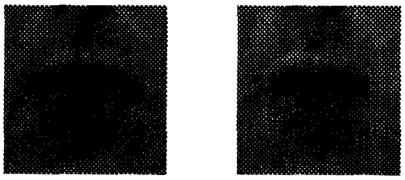

Figure 3 Original and coded mouth images (closed)

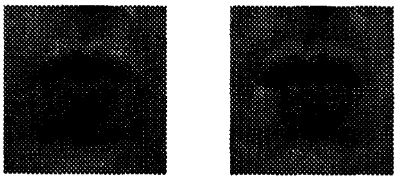

Figure 4. Original and coded mouth image

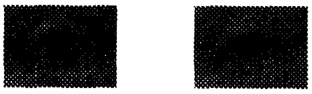

Figure 5. Original and coded eye image(open)