Model-Based Analysis/Synthesis Image Coding with Eye and Mouth Patch Codebooks

Stewart Chao and John Robinson

Abstract

Model-based analysis/synthesis facial coding achieves very low bit rate moving picture transmission. However, the eyes and the mouth are difficult to model reliably and therefore constitute a significant part of the compressed data. We show that use of eye and mouth codebooks significantly reduces this overhead while maintaining reasonable fidelity. Full-motion encoding of a videophone scene is achieved at below 10 kbits/s.

1 Background

Several authors have proposed systems that realize head–and–shoulders moving–picture compression by coding deformations of a wire–frame model shaded with texture from an original neutral picture (for example [1, 2, 3, 4]). Aside from the inherent limitations of using an assumed model for picture content, these schemes still lack good answers to several important questions:

1. What set of parameters is best for describing how the generic model should be conformed to a particular individual's head?
2. How can the conformation of the generic model to a particular individual's head be automated?
3. What set of parameters is best for describing how the model may deform as the person's expression changes?
4. What is the most appropriate way to analyse input pictures in order to extract deformation parameters?
5. What models are appropriate for parts of the body where the 3D wire-frame approach fails (typically the eyes and the mouth)?
6. How should texture and illumination changes be detected and coded?

Our aim is to build a fully automated system, that provides good answers to these questions. As it now exists, our system relies on manual intervention for conforming the model to a particular individual's head, and the parameter set for conformation is therefore provisional. We have previously reported [5] what we believe are strong solutions to the questions of defining and tracking expressions (questions 3 and 4, summarized in section 1.1 below). In this paper we explain how we use 2D pattern patches to address the problem of eyes and mouth coding. This paper is therefore aimed principally at question 5 above.

1.1 Model Based Facial Coding

The model based facial coding scheme introduced by Robinson, Fischl and Miller [5] uses a 3D wire–mesh polygon representation of a generic face, which may be deformed by simulated muscles. The model parameters describe the position and orientation of the head as well as the current level of contraction of each of the muscles being modeled. Manipulation of muscle parameters can provide a very realistic synthesis of human emotions and speech. The model is conformed to specific faces using a non-automated process which produces very accurate models of real faces from 2D images. To provide realistic output, the model is texture–mapped from one original image of the subject that has a neutral facial expression.

A heuristically guided generate-and-test search method is used for determining the parameters of a frame. The algorithm searches for a minimum mean square error between the current frame to be coded and a synthesized face by using the gradient of steepest descent. The search is continued until a minimum is found, then the synthesized frame's parameters are transmitted. Local minima are avoided by incrementing the parameters by large amounts initially, and decreasing to successively smaller ones.

This model can be used for very low bit rate video

telephony in that it captures much of the detail of a human face with a small number of parameters. For each frame, it is only necessary to send updates for parameters which have been modified. The system can code the parameters for moving images at a rate of about 1 kbit/s. It is only necessary to transmit the face's texture information at the start of transmission.

The system has been used to code two short image sequences successfully at a signal–to–noise ratio greater than 18 dB.

The major drawback of the system introduced in [5] is that the teeth, tongue, and eyes are not modeled at all. These areas are represented as patches, and additional texture information must be transmitted for each of them. The amount of bits required by the mouth and eyes are an order of magnitude more than the model parameters.

1.2 Prototype Prediction

Prototype prediction developed by Wollborn [6], was designed as a solution to model failure areas for object-oriented analysis/synthesis coding [7]. Object-oriented analysiss/synthesis coding is a type of unknown-object moving picture coding. Prototype prediction uses previously transmitted similar patterns (image pattern prototypes) to code model failure areas. The difference between the prototype and input area is DCT and DPCM encoded. Prototype prediction reduces the number of bits required to encode an image sequence.

Initially there are no image pattern prototypes at the coder and decoder ends. The color parameters are coded using a special motion compensating hybrid scheme. As images are transmitted, pattern prototypes can be extracted from the images and stored in memory. Subsequent color information can be coded using either a special motion compensating hybrid scheme or prototype prediction, depending on the desired data rate.

This method of using previously transmitted image segments to approximate the blocks of the current frame (of an image sequence) is similar to adaptive vector quantization [8]. However, there are some basic differences. The codebook is generated from previous frames; there are no training sets used. Only the areas that a previous coder (the analysis/synthesis coder) fails to model are encoded, and the patch size and position can change from frame to frame.

2 Eye and Mouth Codebooks

In this section we describe how we have adapted and modified Wollborn's scheme for application to our model-based coder. We define eye and mouth codebooks which store image patches from previous frames in a way analogous to Wollborn's image pattern prototypes. These patches are referenced by the decoder to reconstruct the frame.

2.1 Description

We define a patch as a rectangular extraction of an eye or mouth from the frame. The height and width for each patch is made large enough to contain the whole feature and is rounded up to the nearest multiple of eight. The rounding is to assist JPEG [9] coding of objects during transmission of patches.

At the beginning of an image transmission, there are no eye or mouth patches at the coder and the decoder. The patches from the first frame are extracted and transmitted to the decoder as part of the data to describe the frame. The encoder and decoder codebooks now have one entry each, from the initial frame.

The model-based image analysis process defines the patches of subsequent frames. The patches are extracted and a search is done to find the closest matching patch in the codebook. This is done by finding the lowest mean squared error value of luminance among the codebook entries. If the closest matching mean squared error value is above some (selectable) error threshold, it is assumed there are no entries similar enough to be used to replace the patch. The patch is added to the encoder codebook, and transmitted with the frame information to the decoder (and its codebook) to be used to decode the current frame.

If the match error is below the threshold, the patch is replaced with the closest codebook entry by the decoder. A reference to the codebook entry is transmitted along with the frame information, and the decoder uses the referenced patch from the codebook to resynthesize the face instead of the original object. The replacement patch index and a motion compensation vector are transmitted as side information to the decoder.

When a patch is added to the codebook, the difference between it and the last patch added to the codebook is found. This difference is JPEG encoded and transmitted to the decoder. The decoder JPEG decodes the difference and adds it to the last patch of the codebook. No error residue is transmitted. The resulting sum is added to the codebook.

The codebooks at the coder and decoder side always match.

To compensate inaccuracies in locating the centers of the patches by the analysis/synthesis coder, a motion-compensating vector is used. This is implemented by offsetting the center of the patch by one pixel in any direction. The luminance mean-squared-error calculation is done nine times for each codebook entry.

2.2 Application

Eye and mouth codebooks are being used to improve the performance of the model-based facial coder described in [5]. The un-modeled areas, the mouth and eyes, are coded using patch replacement.

The facial coding scheme is based on an one-layer surface model of the human face. The model assumes that the texture of the human face is constant. The blinking of the eyes and eyelids causes a change in texture which is beyond the scope of the model. A similar problem is the interior of the mouth. Changes in the teeth, tongue, and mouth cavity cause change of texture of the mouth area. Areas where the one-layer facial models fails will therefore be coded using patch codebooks.

The patch replacement processes a frame after the analysis/synthesis has encoded the frame and determined the model failure areas. The facial coding system can automatically extract the eyes and mouth features along with the parameters to model a frame. Problems with accurately determining the centers of the features are addressed by the motion compensation done by the patch selection. The eyes and mouth output is used by the patch replacement system as input.

Each eye and the mouth are treated as separate objects. The patch replacement is done independently on each of the objects. Codebooks are kept for each object.

The patch replacement system outputs a codebook reference and motion compensation vector for each object of each frame. New patches are outputted as JPEG encoded images. The facial coding output along with the eye and mouth coded output is transmitted to the decoder. See figure 1.

The decoder (figure 2) resynthesizes the face using the facial model data first. The eye and mouth codebook reference and motion compensation reference information is decoded and the patches are inserted into the face to complete the image.

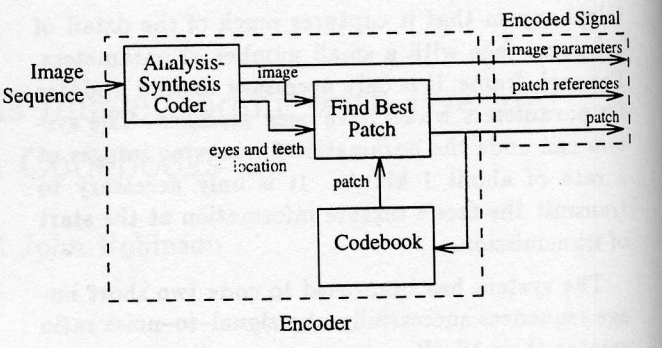

Figure 1: Diagram of encoder

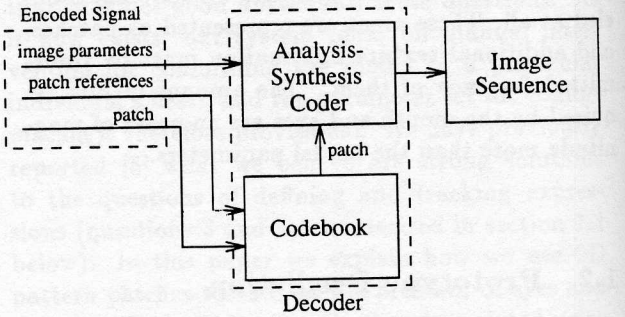

Figure 2: Diagram of decoder

3 Experimental Results

The Miss America test sequence was encoded using the facial coding scheme with patch replacement. The sequence consists of 150 frames. Quality thresholds of 350 mean-squared-error per pel (average squared difference between original and patch pixels) for eyes and 450 for mouth were used. The quality thresholds were found by trial and error. Higher values reduce the data rate at a cost to the image quality. Values lower than 350 and 450 did not improve the subjective image quality significantly. The patch sizes required were 24 by 16 pels for the eyes and 32 by 16 pels for the mouth. Over the whole sequence, the right eye required 11 different patches, the left eye required 15 and the mouth required 33.

3.1 Objective Analysis

The signal-to-noise ratio (figure 3) does not show a change in the fidelity of the sequence since the number of pixels affected by the patch replacement is very small compared to the entire image.

Figure 4 shows the patches used in the sequence. The x-axis represents the frame number and the y-axis represents the frame from which the patch was taken. Points along the diagonal represents patches

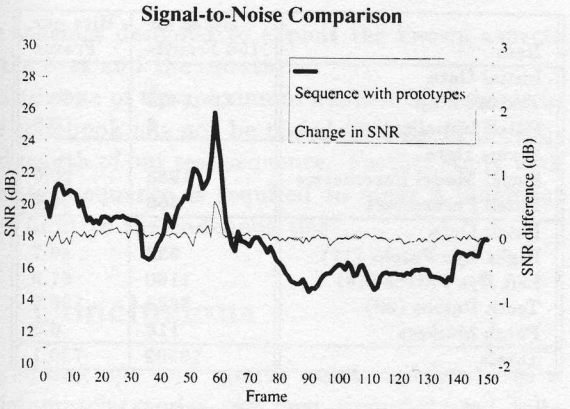

Figure 3: Signal-to-noise ratio of the facial model output using codebooks and the difference between the SNR of the sequence with and without eye and mouth codebooks. Positive values indicate higher SNR for the sequence without codebooks.

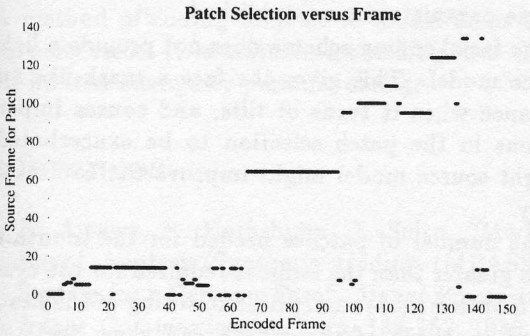

Figure 4: Right eye patch selection used for Miss America sequence.

that are sent to the decoder. The complete codebook for the right eye is shown in figure 5.

3.2 Subjective Analysis

The reproduced sequence has good fidelity but with some noticeable impairments caused by mismatching. Figures 6 and 7 show the output of the model-based system for a particular frame. In figure 6, the eye and mouth patches are filled with texture directly from the input frame; in figure 7, matching has been used to code these areas with previously transmitted patches. For this frame, the mouth substitution is most visible. The mouths of the two frames are noticeably different, however the coded frame still looks very natural.

Single frame analysis of the sequence appears better than frame-by-frame analysis. When the frames are viewed singly the eyes appear natural, how-

Figure 5: The 11 entries in the right eye codebook, from frames 0, 5, 6, 8, 14, 65, 96, 101, 110, 125, and 135.

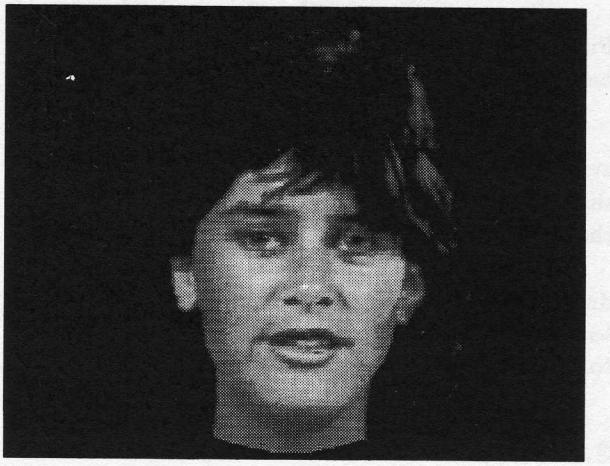

Figure 6: Frame 75 with the eyes and the mouth taken directly from the input frame.

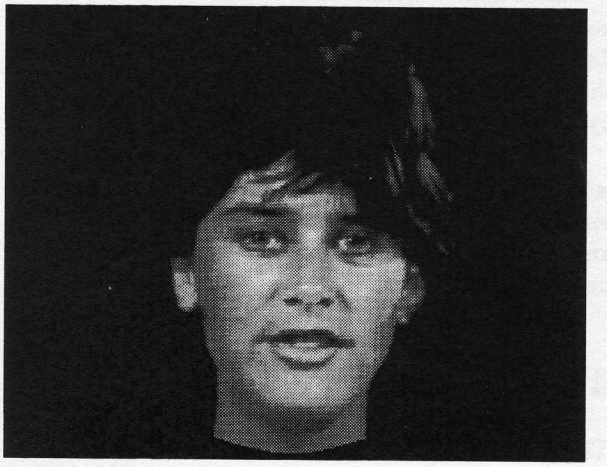

Figure 7: Frame 75 with the eyes and the mouth patches inserted from the codebooks. The right eye comes from frame 65, the left eye from frame 62 and the mouth patch is from frame 71.

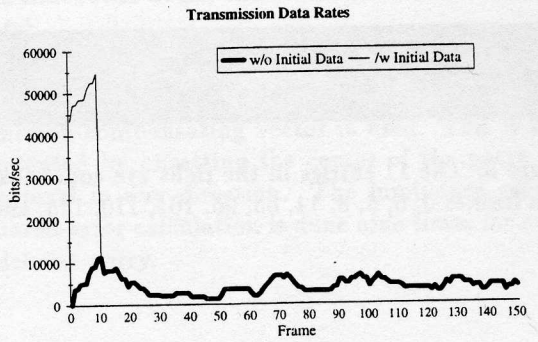

Figure 8: Moving 10 frame data rate. Data rate based on previous 10 frames (1 second).

ever in the continuous sequence the eyes suffer from slight jerkiness caused by the non-fluid motion of the eyes from frame to frame.

The mouth reproduction is smooth. However during parts of the sequence, the tongue is matched to closed lips because of the similarly in color of the tongue and lips.

3.3 Data Rates

Table 1 shows the bytes required to transmit the encoded signal. Overall, not including the initial texture, a rate of 7.2 kbits/sec at 10 frames/sec is achieved. The peak rate for a 10 frame interval is 11.3 kbits/sec.

Each frame requires 90 bits for the facial muscle parameter data, and 26 bits for the patch reference data. Each patch requires 1 byte to tell the receiver the number of bytes to expect for the patch and 1 byte as a stop marker.

Figure 8 shows the data rate for the sequence. It is calculated using the sum of the number of bits sent in the previous 10 frames. The data rate with and without the initial data (texture map for the original "neutral" picture and patch sizes) is shown.

4 Discussion

The sequence used was the 150 frame Miss America sequence. The sequence proved to be too short to take full advantage of the system. We hypothesize that after a certain point of saturation, the codebooks will contain representations of most or all possible eye and mouth states. When the codebook is saturated, no more new patches will need to be transmitted.

Item	Bytes for 150 Frames	Bits per Frame
Initial Data		
Texture Map	5422	289.2
Patch Initialization	6	0.3
Frame Data		
Facial Model Parameters	1688	90.0
Patch Parameters	488	26.0
Patch Data		
Right Eye Patchs (11)	932	49.7
Left Eye Patchs (15)	1160	61.9
Teeth Patchs (33)	3688	196.7
Patch Markers	118	6.3
Totals	13502	720.1

Table 1: Detailed data rate table

The primary drawback to the complete scheme is the time required to generate the muscle parameters. Each frame takes approximately 3 minutes on a DEC 5000/120 to analyze. The patch matching however requires only 1 second per frame once the muscle parameters are established.

The facial coding scheme does not provide a light source model. This gives the face a mask-like appearance when it turns or tilts, and causes imperfections in the patch selection to be exacerbated. A light source model might improve the overall fidelity.

The number of patches needed for the mouth is much greater than the number required for the eyes. The mouth is more susceptible to tilting of the head, and the interior of the mouth can show teeth or empty space. The primary change in the eye is the state of the eyelid.

5 Future Developments

For any moving picture coding system, there are two measures of success: picture fidelity and data rate. Future development of the system reported here will address both of these.

An improved patch selection criterion would enhance picture fidelity. The eyes and the mouth cause different types of fidelity problem. The mean-square-error does not attempt to find the subjective center of the eyes. This causes the decoded eyes to have a "jumpy" or unsteady quality. The mouth has a tendency to match closed-lipped patches with the tongue. Improved patch selection criteria could improve frame-to-frame fidelity of the eyes and the mouth.

After the initial data is transmitted, the patch data forms the bulk of the data transmitted. The data rate can be improved without loss of fidelity by using better coding schemes for the patches or cod-

ing schemes designed to exploit the known aspects of the eyes and the mouth.

The issue of the maximum number of elements in the codebook has not be tested because of the limited length of our test sequence. Further study with a longer sequence is required to make a judgment on a limit to the codebook size.

6 Conclusions

By adding eye and mouth patch codebooks to a facial muscle model, we have demonstrated full–motion head–and–shoulders videphone coding at below 10kbit/s. Although several practical issues still have to be addressed (including model conformation and computational complexity), transmission of good–quality moving pictures over conventional analog telephone lines is possible using this scheme.

A method of coding difficult to model areas has been presented. It is suitable for coding ares that show limited change or periodic change.

References

[1] K. Aizawa, H. Harashima, T. Saito, "Model-Based Analysis/Synthesis Coding (MBASIC) System for a Person's Face", *Signal Processing: Image Communication I*, pp. 139-152, 1989.

[2] H. G. Musmann, M. Hötter, J. Ostermann, "Object-oriented Analysis-Synthesis Coding of Moving Images", *Signal Processing: Image Communication I*, pp. 117-138, 1989.

[3] R. Forchheimer and T. Kronander, "Image Coding - From Waveforms to Animation", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Volume ASSP-37, Number 12, pp. 2008-2023, 1989.

[4] W. J. Welsh, "Model-Based Coding of Videophone Images", *Electronics and Communication Engineering Journal*, Volume 3, Number 1, pp. 29-36, 1991.

[5] J. Robinson, J. Fischl, B. Miller, "Parameter Tracking in a Muscle-Based Analysis/Synthesis Coding System", *Proceedings of the 1993 Picture Coding Symposium*, pp. 2.3.1-2.3.2, March 1993.

[6] M. Wollborn, "Object-Oriented Analysis-Synthesis Coding Using Prototype Prediction for Colour Update", *Proceedings of the 1993 Picture Coding Symposium*, pp. 17.6.1-17.6.2, March 1993.

[7] M. Hötter, "Object-Oriented Analysis-Synthesis Coding of Moving Two-Dimensional Objects", *Image Communication*, Volume 2, Number 4, December 1990, pp. 409-428.

[8] N. M. Nasrabadi, "Image Coding Using Vector Quantization: A Review", *IEEE Transactions on Communications*, Volume 26, Number 8, August 1988, pp. 957-971.

[9] G. K. Wallace, "The JPEG Still Picture Compression Standard", *Communications of the ACM*, Volume 34, Number 4, April 1991, pp. 30-44.