WIRE FRAME FITTING FOR AUTOMATIC TRACKING IN MODEL-BASED VIDEO CODING

P. M. Antoszczyszyn, J. M. Hannah, P. M. Grant

Department of Electronics and Electrical Engineering,
The University of Edinburgh, King's Buildings,
Edinburgh, EH9 3JL, Scotland, UK

e-mail: plma@ee.ed.ac.uk

ABSTRACT

Model-based video coding requires the application of both image processing and machine vision techniques for proper fitting of the semantic model and its subsequent tracking throughout the rest of the sequence of a certain type (e.g. 'head-and-shoulders' or 'head-only'). In this article a method of automatic semantic wire-frame fitting based on a reference data-base of facial images is presented. The method has been tested on a widely used data-base of images with very good results. It was possible to accurately retrieve the position of the facial features in all cases. The position of the facial features in initial frames can subsequently be used in automatic tracking. Results of automatic fitting are presented as a part of this contribution. Experimental results are also available on-line in the form of compressed movies from our Internet site at http://www.ee.ed.ac.uk/~plma/.

1. INTRODUCTION

Application of traditional (block-based) moving image coding techniques in transmission channels of extremely low data-rate (below 10 kbit/s) results in unacceptable artefacts. Model-based techniques offer an alternative approach to the problem of transmission of video in extremely low data-rate environments (e.g. mobile communication, PSTN lines in certain countries) where the approximate content of the video scene is known.

Despite the introduction of other moving image coding techniques based on vector quantisation [1], fractal theory [2] and wavelet analysis [3] it is still not possible to send video over extremely low bit-rate channels with acceptable quality. A very promising approach using scene analysis techniques was proposed by Musmann *et al.* [4]. However, according to the assessments of Aizawa *et al.* [5] and Forchheimer [6] only the application of semantic model-based techniques offers the potential to obtain data-rates below 10 kbit/s for *head-and-shoulders* video sequences.

The concept of model-based communication can be briefly explained in the following way. A semantic model of the scene is shared by the transmitter and the receiver. (Since our main concern is a typical videophone scene -'head-and-shoulders' or 'head-only' - the *Candide* wire frame model [6] was used - Figure 1). With each subsequent frame of the video sequence the position of the vertices of the wire-frame are automatically tracked. The initial and subsequent positions of the wire-frame are transmitted in the form of 3D co-ordinates over the low bit-rate channel along with the texture of the face from the initial frame of the sequence. Knowing the texture of the scene from the initial frame and the 3D positions of the vertices of the wire-frame in subsequent frames it is possible to reconstruct the entire sequence by mapping at locations indicated by the transmitted vertices.

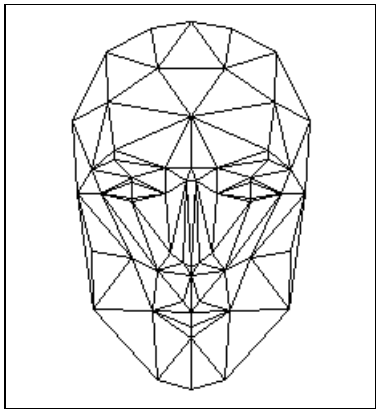

Figure 1. Candide wire-frame model of face

However, before the transmission commences, the wire-frame must be mapped onto the actual image of the subject (the automatic fitting problem). Once the fitting is completed successfully, the co-ordinates of the wire-frame must be updated on frame-by-frame basis (the automatic tracking problem).

The automatic wire-frame fitting method proposed by Welsh [7] utilises the idea of 'snakes' (active contours). Different approaches were presented by Reinders *et al.* [8] and Seferidis [9].

In this article we present developments of the method we recently proposed [10] including the results of tests carried out on a wider range of images.

2. FITTING ALGORITHM

Our approach is based on the principal components analysis (PCA) of a code-book of facial images. In this case, the MIT facial images were used. In the first step of the PCA, the

eigenvectors of the covariance matrix **S** of the sequence **X** of *M*, *N* - dimensional input column vectors: $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ ...\ \mathbf{x}_M]$, $\mathbf{x}_j = [x_{ji}]$, $i = 1..N$, $j = 1..M$, must be found. In our experiments the sequence **X** consists of *M* sub-images extracted from *M* different images of the facial code-book. In our analysis we utilise five types of sub-image sequences: one sequence of sub-images containing faces, and four sequences of sub-images containing important facial features: the left eye, the right eye, the nose and the lips. The dimensions of all sub-images in a particular sequence are the same. However, the dimensions of a sub-image containing one feature (e.g. the left eye) may be different from those containing another feature (e.g. lips). The analysis of each sequence is performed in the same way regardless of the size of the sub-image in the sequence.

Each sub-image containing a facial feature is first histogram-equalised. The sequence of such pre-processed sub-images is converted into 1D column vectors $\mathbf{x}_j$ by scanning the image line by line. An image consisting of *R* rows and *C* columns would therefore produce a column input vector consisting of $N = C \times R$ rows. We obtain the covariance matrix from the following relationship:

$$\mathbf{S} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2\ ...\ \mathbf{y}_M]$, $\mathbf{y}_j = \mathbf{x}_j - \mathbf{m}_x$ and $\mathbf{m}_x$ is the expected value of the sequence **X**. We can find the *i*-th principal component z_i of the initial set from the following equation:

$$z_i = \mathbf{u}_i^{\mathrm{T}} (\mathbf{x}_i - \mathbf{m}_x) \tag{2}$$

where $\mathbf{u}_i$ is the *i*-th eigenvector of the covariance matrix **S.** Even for small images, the size of the covariance matrix can be too large to handle by common computing equipment (e.g. a sequence of images consisting of 100 columns and 100 rows would result in a $100^2 \times 100^2$ covariance matrix). However, if the number of images *M* in the sequence **X** is considerably smaller than the dimensions of the images themselves ($N = C \times R$), the above problem can be overcome (Murakami and Kumar [11]). The eigenvectors of the covariance matrix $\mathbf{S} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ can be expressed as a linear combination of eigenvectors of the matrix $\mathbf{C} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}$. Since matrix **C** is $M \times M$, the computational costs of finding the eigenvectors of the matrix **S** are greatly reduced (in our research $M < 20$ and $N < 100$ thus the problem is reduced to calculations involving matrices smaller than 20×20).

The same algorithm is applied to each sub-image sequence. We therefore obtain five principal component spaces: one for each analysed sequence of histogram-equalised sub-images. At the same time all the images for the facial code-book are manually pre-fitted with the *Candide* (Figure 1) wire-frame. Both processes (calculation of principal components spaces and wire-frame pre-fitting) are parts of the system preparation and do not influence the speed of the actual fitting algorithm in any way.

On-line processing (automatic wire-frame fitting) starts with the unknown (incoming) image. The fitting is performed in two stages. In the first stage (the coarse stage) the approximate position of the subject's face is established. This is accomplished by the analysis of the principal components space of the sequence of sub-images containing faces (the face sequence) extracted from the facial code-book. The analysis in the coarse stage is performed as follows. A sub-image of the same dimensions as the images from the face sequence is extracted at every possible location in the unknown image. It is subsequently histogram-equalised and converted into a 1D column vector by scanning the image line by line, and then projected onto the principal components space of the face sequence. For this purpose we use equation (2) with a single modification: the image $\mathbf{x}_i$ is now a sub-image extracted at the *i*th position on the unknown image, not an image from the face sequence. Since the reference principal components space was created using face sub-images only, the projection of an unknown image tells us how similar the unknown sub-image is to all the images from the face sequence. Or, in other words, it allows us to judge whether the analysed (unknown) sub-image is a face or not. This can be quantitatively described by the following distance measure:

$$d_i = \left\| \mathbf{y}_i - \mathbf{r}_i \right\| \tag{3}$$

Where the $\mathbf{r}_i$ represents the reconstruction of the *i*-th image ($\mathbf{x}_i$) after its projection onto the principal components space of the face sequence. The distance (3) is calculated for each sub-image extracted from the unknown image. The spatial location at which the distance (3) reaches a minimum is the approximate position of the face on the unknown image. This is the best match location. Once the coarse position of the object's face is estimated, we still have to choose the wire-frame that should be fitted to the unknown image. The distance (3) does not refer to any specific sub-image from the face sequence. In order to find out which wire-frame is most appropriate at the best match location we use the following distance measure:

$$d_j = \left\| \mathbf{a} - \mathbf{b}_j \right\|^2 \qquad j = 1,..,M \tag{4}$$

where **a** is the projection of the sub-image extracted from the unknown image at the best match location and $\mathbf{b}_j$ is the projection of the *j*-th image from the face sequence onto principal components space. This distance measure was proposed for facial recognition purposes by Turk and Pentland [12]. Again the minimum distance (4) tells us which wire-frame to use (index *j*) for coarse fit at the best match location. Once both the best match location and the wire-frame are established, the first stage (the coarse stage) of our algorithm is concluded.

In the second stage of the algorithm both the results from the first stage and the information about the geometry of the human face are combined in order to achieve faster and more reliable operation.

We now analyse the principal components space of the sequence of sub-images containing important facial features (the left eye sequence, the right eye sequence, the nose sequence and the lips sequence) previously extracted from the facial code-book. The method of accurate fitting in stage two will be explained using the example of the left eye sequence. A sub-image of the same dimensions as the images from the left eye sequence is extracted at every possible location of a search region. This search region is centred on the coarse position of the pupil of the left eye (as

indicated after completion of the first stage). The dimensions of the search region are adjustable, but twice the size of the sub-image seems to give good results. The further analysis is similar to that carried out in the coarse stage: we determine the best match location of the left eye of the subject using distance (3). This is followed by a search for the most appropriate wire-frame of the left eye using distance (4). The vertices corresponding to the left eye (Figure 2) are subsequently extracted from this wire-frame and fitted (at the best match location of the left eye) to the wire-frame chosen in the coarse stage. The same algorithm is repeated for the right eye, the nose and the lips.

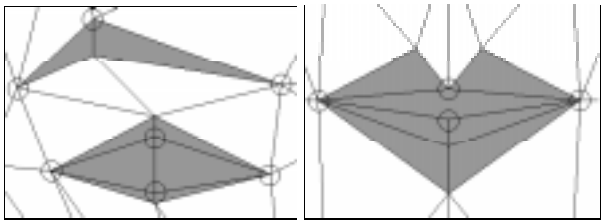

Figure 2: The left eye (left) and the lips (right) vertices

Thus the final wire-frame is assembled of five partial wire-frames: the wire-frame of the face (found in the first stage), and four wire-frames of the important facial features (found in the second stage). This concludes the automatic wire-frame fitting algorithm.

3. EXPERIMENTAL RESULTS

We have used the images from the MIT facial data-base in order to create principal component spaces of faces and features and images from the Manchester facial data-base for automatic fitting tests. The coarse fitting alone gave positive results in more than 85% of cases. This constitutes a significant improvement over results we achieved without the application of histogram equalisation on the same set of images. A quantitative assessment of the accuracy of the fit using PSNR or similar measures seems inappropriate in our case. Small PSNR values can be misleading if the automatic fit is inaccurate in the area occupied by facial features (especially the lips and the eyes).

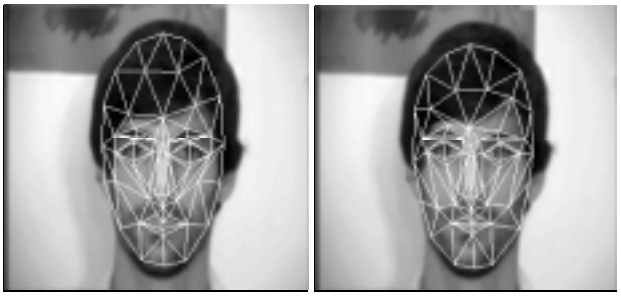

Figure 3: Manual (left) and automatic (right) fit

On the other hand, relatively large values of PSNR may result in very little subjective degradation in the quality of the animated picture.

Therefore, in order to assess the accuracy of the fitting algorithm, we decided to perform subjective tests. Each analysed image was fitted both automatically (using our algorithm) and manually (Figure 3). Subsequently the same action units (Ekman and Friesen [13]) were changed in both models and the fidelity of reconstruction of facial expression was judged subjectively by creating short movies with the face fitted manually on the left and the same face fitted automatically on the right. Figures 4-9 present some stills from these short movies (manually fitted face on the left, automatically fitted face on the right). The full compressed movies are available from our Internet site at http://www.ee.ed.ac.uk/~plma/. The action units corresponding to the presented stills are listed in Table 1. The stills show that the method works reliably and clearly demonstrates the successful operation of our algorithm. It can be seen, that the reconstruction of facial expressions gives excellent results, even in the case of the most difficult action unit: lips open-close (Figures 4-9). This is due to very good fit of the centre of the lips of the wire-frame to the centre of the lips of the subject.

Action Unit	Level	Figure
Lip corner depressor	+0.3	4
Brow lowerer	+0.3	5
Lip stretcher	+0.5	6
Lip stretcher	-0.3	7
Upper lip raiser	+0.1	8
Lower lip lowerer	+0.5	9

Table 1: Action units tested

4. CONCLUSIONS

We have proposed an improved algorithm for automatic wire-frame fitting for model-base moving image coding. We have created short movies with the face fitted manually on the left and the same face fitted automatically on the right in order to judge the reliability of the algorithm. The movies were created for all tested facial images from the Manchester data-base. We have shown that the images fitted automatically and manually can be animated to give very similar subjective results by changing action units of the *Candide* wire-frame model. The wire-frame fit in the areas occupied by the important facial features (the left and right eye, the lips, the nose) was almost perfect. This is of particular importance for the lips and the eyes whereas imperfections in the fit in other facial areas are more tolerable. In our future research we intend to apply more detailed wire-frame models, especially in the areas occupied by the lips and the eyes.

5. ACKNOWLEDGMENT

Paul Antoszczyszyn acknowledges the support of The University of Edinburgh through a Postgraduate Studentship.

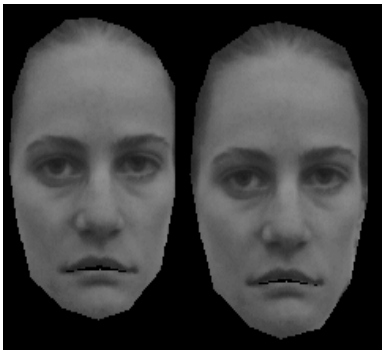

Figure 4

Figure 5

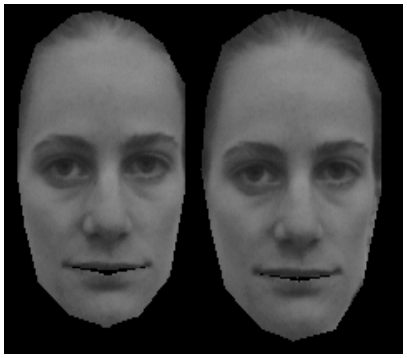

Figure 6

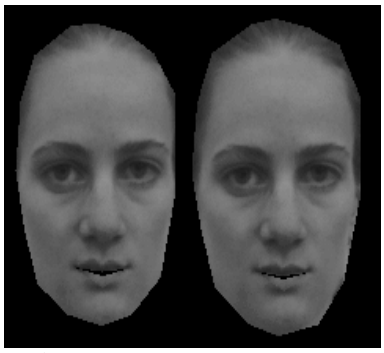

Figure 7

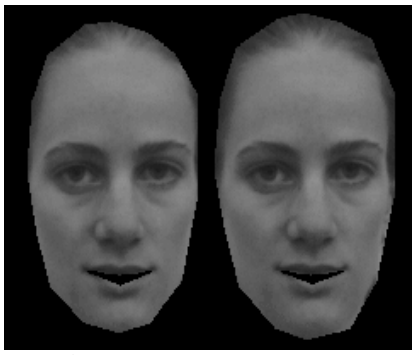

Figure 8

Figure 9

6. REFERENCES

[1] Gersho A. "On the structure of vector quantizers". *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 157-166, March 1982.

[2] Jacquin A.E. "Image coding based on a fractal theory of iterated contractive image transformations". *IEEE Transactions on Image Processing*, vol. 1, no. 1, pp. 18-30, January 1992.

[3] Antonini M., Barlaud, M., Mathieu P. and Daubechies, I.: 'Image coding using wavelet transform', *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 205-220, April 1992.

[4] Musmann H.G., Hoetter M. and Ostermann J. "Object-oriented analysis-synthesis coding of moving images". *Signal Processing: Image Communication*, vol. 1, no. 2, pp. 117-138, October 1989.

[5] Aizawa K., Harashima H. and Saito T. "Model-based analysis synthesis image coding (MBASIC) system for a person's face". *Signal Processing: Image Communication*, vol. 1, no. 2, pp. 139-152, October 1989.

[6] Forchheimer R. and Kronander T. "Image coding - from waveforms to animation" *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 37, no. 12, pp. 2008-2023, December 1989.

[7] Welsh B. "Model-based coding of video images". *Electronics and Comms. Eng. Journal*, vol. 3, no. 1, pp. 29-38, February 1991.

[8] Reinders M.J.T., van Beek P.J.L., Sankur B. and van der Lubbe J.C.A. "Facial feature localization and adaptation of a generic face model for model-based coding". *Signal Processing: Image communication*, vol. 7, no. 1, pp. 57-74, March 1995.

[9] Seferidis V. "Facial feature estimation for model-based coding". *Electronics Letters*, vol. 27, no. 24, pp. 2226 - 2228, November 1991.

[10] Antoszczyszyn P.M., Hannah J.M. and Grant, P.M. "Facial features model fitting in semantic-based scene analysis". *Electronics Letters*, vol. 33, no, 10, pp. 855-857, May 1997.

[11] Murakami H. and Kumar V. "Efficient calculation of primary images from a set of images" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 5, pp. 511-515, September 1982.

[12] Turk M. and Pentland A. "Eigenfaces for recognition". *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, Winter 1991.

[13] Ekman P. and Friesen W.V. "Facial action coding system", Consulting Psychologists Press Inc., Palo Alto, California, 1977.