FACIAL FEATURES MOTION ANALYSIS FOR WIRE-FRAME TRACKING IN MODEL-BASED MOVING IMAGE CODING

Paul M. Antoszczyszyn, John M. Hannah and Peter M. Grant

Department of Electrical Engineering,
The University of Edinburgh, EH9 3JL Edinburgh, UK
plma@ee.ed.ac.uk

ABSTRACT

This paper addresses the problem of wire-frame tracking by accurate analysis of the motion and the shape of the facial features in head-and-shoulders scenes. Accurate wire-frame tracking is of paramount importance for correct reconstruction of the encoded image, especially in the areas occupied by the lips and the eyes. An entirely new algorithm for tracking the motion of a semantic wire-frame (Candide) by analysis of the principal components of sub-images containing important facial features of the speaker's face is proposed. This algorithm is suitable for tracking both global motion (motion of the speaker's head) and local motion (motion of the facial features). The algorithm was tested on numerous head-and-shoulders sequences with excellent results.

1. INTRODUCTION

Aizawa *et al.* [1], and Forchheimer and Kronander [2] suggest that moving image compression techniques based on semantic models are capable of achieving data-rates below 10 kbit/s. This would allow real-time video communication over PSTN lines. The two main problems in model-based moving image coding are automatic wire-frame fitting and automatic wire-frame tracking.

Tracking algorithms for semantic-based moving image coding have been proposed by Li and Forchheimer [3] who employed optical flow analysis and Kokuer and Clark [4] who adapted a correlation approach. Our method is based on the analysis of the principal components of a set of images and imposes no restrictions on the amount of motion in the scene.

2. EXPERIMENTAL METHOD

We have concentrated our efforts on tracking the *motion* and the *shape* of the left eye, the right eye, the nose and the lips. These facial features will be referred to as the *important facial features*. Each important facial feature is tracked separately thus its 2D co-ordinates can be used to determine the current position of the speakers head. The algorithm is identical for each important facial feature, but will be described for the left eye.

Our tracking method is based on principal components analysis (PCA) and we believe this to be the first attempt to utilise PCA for motion analysis in model-based coding. In the first step of the PCA, the eigenvectors of the covariance matrix $\mathbf{S}$ of the sequence $\mathbf{X}$ 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 analysis the input sequence consists of sub-images containing the left eye of the speaker extracted from M initial frames of the test sequence (Figure 1). The input sequence of sub-images (further referred to as the *initial set*) 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:

$$\mathbf{S} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} \qquad (1)$$

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

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

where $\mathbf{u}_i$ is the i-th eigenvector of the covariance matrix $\mathbf{S}$. 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 50 columns and 50 rows would result in a $50^2 \times 50^2$ covariance matrix). However, if the number of images M in the sequence $\mathbf{X}$ is considerably smaller than the dimensions of the images themselves ($N = C \times R$), the above problem can be overcome. According to the method of singular value decomposition (SVD) [5], the eigenvectors of the covariance matrix $\mathbf{S} = \mathbf{YY}^T$ can be expressed as a linear combinations of eigenvectors of a matrix $\mathbf{C} = \mathbf{Y}^T\mathbf{Y}$. Since matrix $\mathbf{C}$ is $M \times M$, the computational costs of finding the eigenvectors of the matrix S are greatly reduced. In our research $M < 20$ and $N < 50$. Thus the problem is reduced to calculations involving matrices smaller than 20 × 20.

Once the eigenvectors of the covariance matrix **S** of the initial set of M sub-images containing the left eye of the speaker extracted from the M initial frames of the sequence are calculated, the automatic tracking commences with frame M +1. The initial position of the left eye in frame M + 1 (current frame) is assumed to be the same as in frame M (previous frame). This view is subsequently verified in the following way. The sub-images within the search range centred on the initial position of the left eye in the M +1-th frame are extracted from the current frame (e.g. for a search range of 15 × 15 we obtain a set of 225 images). These images are referred to as the *extracted set* (Figure 1). It is the task of the algorithm to find the *best match* image among the images from the extracted set.

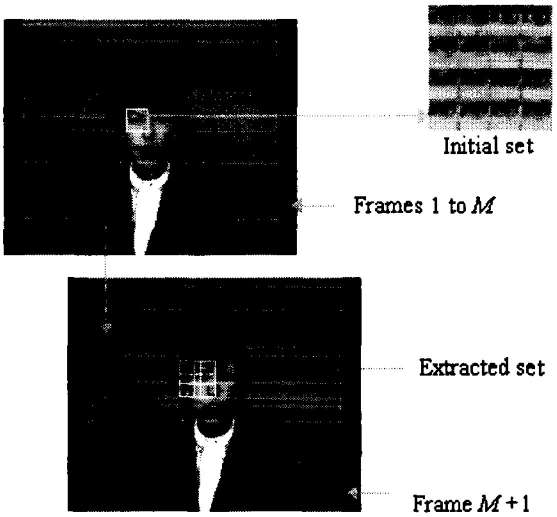

Figure 1: Automatic tracking system

The dimensions of the images from the extracted set are identical to those from the initial set. Since the images from the extracted set are *similar* to those from the initial set we can assume, that they can be projected onto the principal components space created by input vector **X** of the initial set of images. For this purpose we use equation (2) with a single modification: the image $\mathbf{x}_i$ is now the i-th image from the extracted set, not the initial set. If we transform this image using the principal components space created by the input vector **X** we will obtain certain image $\mathbf{r}_i$. Since the principal components space was created using the images from the *initial set* (1), the Euclidean distance between the $\mathbf{x}_i$ image and the $\mathbf{r}_i$ image will tell us how similar the $\mathbf{x}_i$ image from the *extracted set* is to all the images from the *initial* set:

$$d_i = \| \mathbf{x}_i - \mathbf{r}_i \| \tag{3}$$

The i-th image from the extracted set for which the distance d_i is minimal, is the best match image. The co-ordinates of its centre on the M +1-th frame are the co-ordinates of the left eye of the speaker on the M +1-th frame. This algorithm is repeated for the remaining frames of the test sequence and is identical for the remaining important facial features: the right eye, the lips and the nose. The distance measure (3) was first proposed by Turk and Pentland [6].

3. EXPERIMENTAL RESULTS

We have tested our automatic tracking algorithm on numerous commonly used *head-and-shoulders* video sequences: *Miss America* (352 × 240 pixels, 150 frames), *Claire* (360 × 288 pixels, 168 frames), *Car Phone* (176 × 144 pixels, 400 frames), *Grandma* (176 × 144 pixels, 768 frames), *Salesman* (360 × 288 pixels, 400 frames) and *Trevor* (256 × 256 pixels, 100 frames). The track of *all* facial features was maintained for *all* tested sequences.

Figure 2: *Miss America*

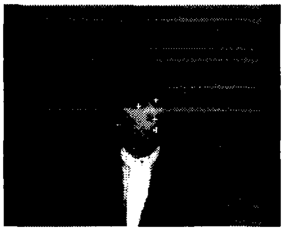

Figure 3: *Claire*

Figure 4: *Grandma*

Figure 5: *Car phone*

Figure 6: *Salesman*

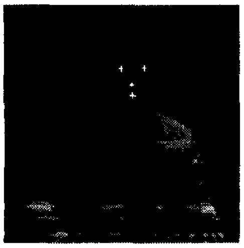

Figure 7: *Trevor*

The track was maintained even when the facial features were partially occluded by the speaker's hand (*Salesman*) or when they radically changed shape (eye close-open, mouth close-open).

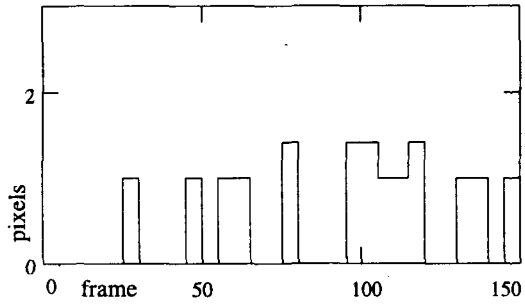

Figure 8: *Miss America*: The lips tracking error profile

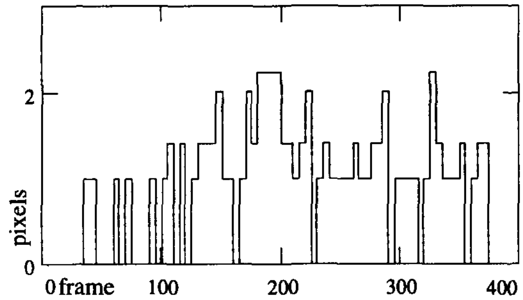

Figure 9: *Car Phone*: The left eye tracking error profile

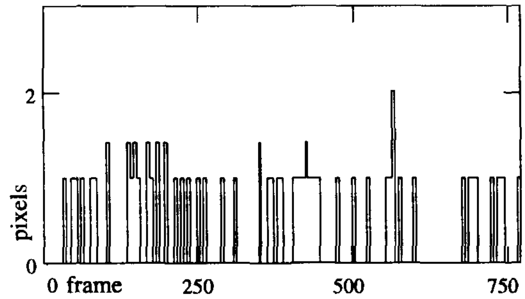

Figure 10: *Grandma*: The nose tracking error profile

Also tracking of the eyes of the subjects wearing glasses (*Grandma*, *Trevor*) was completely successful. The sequences contained moderate zoom, rotation and translation. We have created short movies with white crosses centred on the important facial features for all test sequences (Figures 2 to 7). These crosses were observed to track the features with remarkable precision.

However in order to assess the accuracy of the tracking algorithm, the 2-D positions of the important facial features were extracted manually from every fifth frame of the test sequences. The Euclidean distance between the feature tracked manually and automatically (on every fifth frame) is a good measure of the accuracy of the tracking method. Typical error distance profiles for some of the facial features are presented (Figures 8 to 10).

Facial feature	Mean error [pixels]	Standard deviation [pixels]
Left eye	0.6	0.7
Right eye	0.8	0.7
Nose	0.6	0.6
Lips	0.5	0.6

Table 1: Tracking results for *Miss America*

Facial feature	Mean error [pixels]	Standard deviation [pixels]
Left eye	0.4	0.5
Right eye	0.6	0.7
Nose	0.8	0.6
Lips	1.0	0.6

Table 2: Tracking results for *Claire*

Facial feature	Mean error [pixels]	Standard deviation [pixels]
Left eye	0.8	0.8
Right eye	0.8	0.9
Nose	0.9	0.6
Lips	0.7	0.7

Table 3: Tracking results for *Trevor*

We have also calculated the mean error and standard deviation for all the tracked facial features. A few typical results are presented in Tables 1 to 3). As can be seen, the mean error for all the facial features in all the sequences was no more than 1 pixel.

Since we wish to utilise the *Candide* [2] wire-frame, in order to reconstruct the local motion (e.g. lips close-open, eyes close-open) we must be able to track reliably the motion of the vertices assigned to the selected facial features (Figure 11). We utilised the same algorithm, but this time the initial set images were centred on the points of the image that corresponded to the positions of the wire-frame vertices of a particular facial feature. Again observation of test video sequences re-created from the results of the algorithm showed an excellent tracking performance.

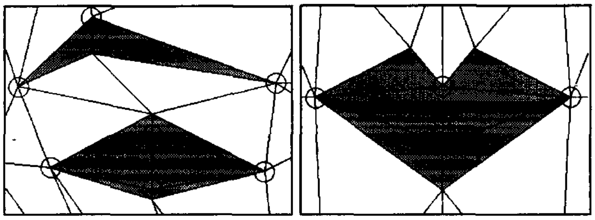

Figure 11: Tracking vertices (in circles) of the left eye (left) and the lips (right)

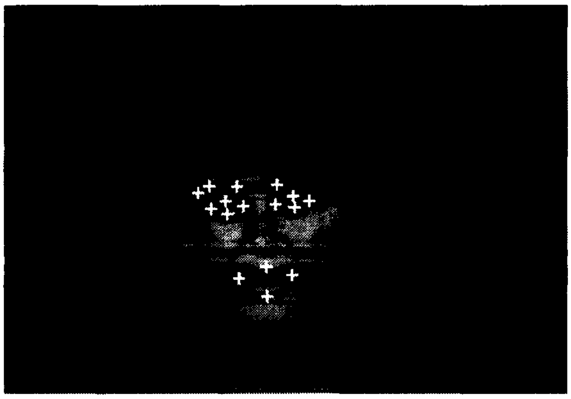

Figure 12: Tracking of the shape of the facial features

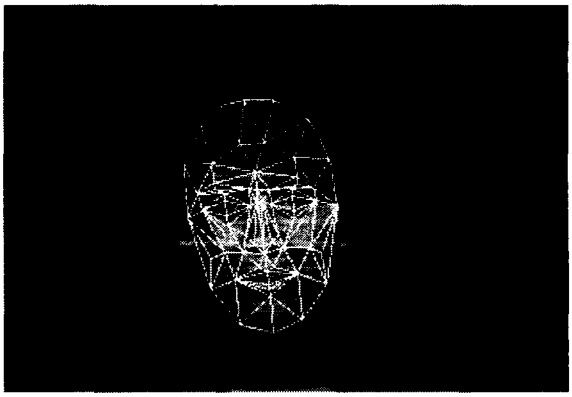

Figure 13: Manipulating the *Candide* wire-frame

The tracked vertices were subsequently used as anchors for vertices of the *Candide* wire-frame model (Figures 12 and 13). Thus the wire-frame model was driven by the global motion of the speaker's head and local motion of the facial features.

4. CONCLUSIONS

We have developed a new and reliable algorithm for automatically tracking the motion of facial features in *head-and-shoulders* scenes. The algorithm is based on eigenvalue decomposition of sub-images containing important facial features: the eyes, the nose and the lips. The algorithm was tested on numerous sequences containing limited pan, rotation and zoom of the speaker's head, with excellent results. Since all the facial features are tracked independently, the algorithm could be easily adapted for use on a parallel processing system.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

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

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

[3] H. Li and R. Forchheimer, "Two-view facial movement estimation", IEEE Transactions on Circuits and Systems for Video Technology, vol. 4, no. 3, pp. 276-287, June 1994.

[4] M. Kokuer and A. F. Clark, "Feature and model tracking for model-based coding", Proceedings of 1992 IEE International Conference on Image Processing and Its Applications, Maastricht, Netherlands, 7-9 April 1992, pp. 135-138.

[5] H. Murakami and V. Kumar, "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.

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