AUTOMATIC FITTING AND TRACKING OF FACIAL FEATURES IN HEAD-AND-SHOULDERS SEQUENCES

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

The University of Edinburgh
Department of Electrical Engineering
Edinburgh, Scotland, 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'). A method of automatic semantic wire-frame fitting and tracking based on principal component analysis using an independent reference data-base of facial images is presented. The method has been tested on widely used 'head-and-shoulders' video sequences with very good results. It was possible to accurately retrieve the position of the desired facial features in all cases. The position of the facial features in initial frames was subsequently used in automatic tracking. Experimental results are presented as a part of this contribution. Compressed movies illustrating these results can be viewed from our Internet site http://www.ee.ed.ac.uk/~plma/.

1. INTRODUCTION

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 artifacts. Model-based techniques offer an alternative approach to the problem of transmission of moving images 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 quantization [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, it seems that only the application of semantic-based techniques will potentially allow transmission of moving images over transmission media with extremely narrow bandwidth. According to the work of Aizawa *et al.* [5] and Forchheimer [6] it is possible 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). The model must be initially pre-fitted to the actual scene without human intervention (the fitting problem). With each subsequent frame of the video sequence the position of the vertices on the transmitter side of the wire-frame must be tracked reliably (the tracking problem). The initial and subsequent positions of the wire-frame are transmitted in the form of 3D coordinates 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 and the 3D positions of the vertices of the wire-frame in subsequent frames it is possible to reconstruct the entire sequence by mapping the texture of the initial frame of the sequence at locations indicated by the transmitted vertices.

In earlier contributions we have proposed separate solutions to the fitting [7] and tracking [8] problems. Here we propose a unified approach to solving both facial feature fitting and tracking issues by the application of principal component analysis to a sequence of sub-images for automatic retrieval and subsequent automatic tracking.

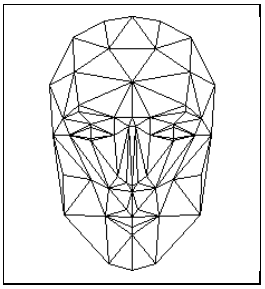

Figure 1. Candide wire-frame model of face

2. ALGORITHM DESCRIPTION

2.1 Automatic fitting

Our approach is based on the principal component analysis (PCA) of a sequence of sub-images extracted from a facial code-book. From each image in the data-base a sub-image containing the subject's face and the feature to be tracked (e.g. the left eye) is extracted (Fig. 2). In this way, two separate sequences, one containing faces, the other the feature to be tracked, are obtained. Each image in a sub-image sequence is first submitted to histogram equalization and then converted into a 1D column vector $\mathbf{x}_i$ (by line-by-line scanning). The ith principal component z_i of the sub-image sequence can be found from the following equation:

$$z_i = \mathbf{u}_i^T (\mathbf{x}_i - \mathbf{m}_x), \quad (1)$$

where $\mathbf{u}_i$ is the ith eigenvector of the covariance matrix $\mathbf{S} = \mathbf{Y}\mathbf{Y}^T$, $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2\ ...\ \mathbf{y}_M]$, $\mathbf{y}_j = \mathbf{x}_j - \mathbf{m}_x$ and $\mathbf{m}_x$ is the expected value of the sequence of images. We obtain separate principal component spaces: one for each analyzed sequence of sub-images.

The above process is not a part of the automatic fitting-tracking method. It is performed off-line and does not influence the speed of the on-line algorithm in any way, i.e. the principal component spaces created at this stage are known before the on-line algorithm commences. In our research we used sub-image sequences extracted from an MIT facial data-base [9].

The on-line algorithm commences with automatic retrieval of the facial features from the initial frames of the sequence. This process is performed in two stages. In the first stage the approximate position of the subject's face is established by analysis of the principal component space of the sequence of sub-images containing faces (the face sequence). A sub-image of the same dimensions as the images from the face sequence is extracted at every possible location in the unknown image and projected onto the principal component space of the face sequence. For this purpose we use equation (1) with a single modification: the image $\mathbf{x}_i$ is now a sub-image extracted at the ith position in the unknown image, not an image from the reference sequence of faces. How similar the unknown sub-image $\mathbf{x}_i$ and the images from the face sequence are can be quantitatively described by the following distance measure:

$$d_i = \| \mathbf{y}_i - \mathbf{r}_i \| \quad (2)$$

where the $\mathbf{r}_i$ represents the reconstruction of the ith image ($\mathbf{x}_i$) after its projection onto principal component space of the face sequence. The extraction point of the sub-image $\mathbf{x}_i$ for which d_i reaches a minimum is the 'best match location' - the coarse position of the face. Such a distance measure has previously been proposed for facial recognition purposes [10].

In the second stage we analyze the principal component space of the sequence of sub-images containing the facial feature to be tracked extracted from the facial code-book (the feature sequence - Fig. 2). A sub-image of the same dimensions as the images from the feature sequence is extracted at every possible location in a search region. The search region is centered on the coarse position of a particular feature (e.g.: the pupil of the left eye - as indicated from completion of the coarse face fit stage). 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 (2).

The above two-stage algorithm is repeated for M initial frames of the analyzed sequence (in our case $M = 16$). Once the position of the feature is retrieved from the initial M frames, the principal component spaces created using images from the independent facial data-base are abandoned. They are replaced by a new set of principal component spaces created using the positions of the speaker's head and facial features retrieved automatically from the initial M frames of the video sequence. This newly created set of principal component spaces is utilized in the tracking algorithm and is not updated throughout the tracking process.

2.2 Automatic tracking

Once the new set of principal component spaces is created, the algorithm switches to tracking mode. The initial position of the facial feature in the next (M+1) frame of the video sequence is assumed to be the same as in the frame on which the last automatic feature retrieval was performed (frame M). This is subsequently verified in the following way. Sub-images within the search range are extracted from the next frame. Each sub-image is submitted to histogram equalization and then projected onto the principal component space of the desired facial feature using equation (1). The sub-image for which distance (2) is minimized is assumed to hold the searched-for facial feature, and its 2-D position is the position of facial feature in the next frame. The above algorithm is repeated on a frame-by-frame basis until the end of the video sequence.

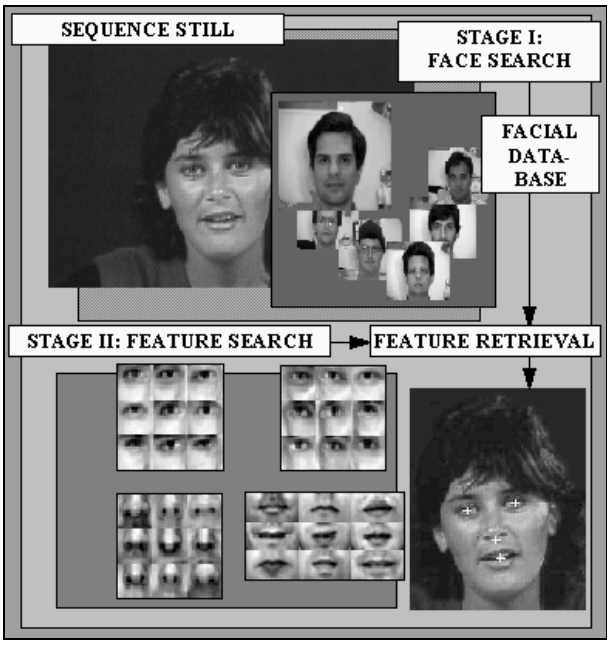

Figure 2. Candide wire-frame model of face

3. EXPERIMENTAL RESULTS

The algorithm was tested on a range of widely used head-and-shoulders sequences and was found to perform very well. Tracking of all facial features was maintained in all sequences, as observed from short movies created showing white crosses centered on the important facial features. Some stills from these movies are presented in Figures 3-7. Full compressed movies showing the results are also available from our Internet site at http://www.ee.ed.ac.uk/~plma/.

Tracking 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). Also tracking of the eyes of the subjects wearing glasses (*Grandma*, *Trevor*) was successful. These test sequences included moderate zoom, rotation and translation.

Figure 3. *Miss America* video still

Figure 4. *Claire* video still (352 x 240 pixels)

Figure 5. *Salesman* video still

Figure 6. *Car Phone* (left) and *Grandma* (right) stills

Figure 7. *Trevor* video still

In order to quantitatively describe the accuracy of our method, the 2-D positions of the important facial features were extracted manually from every fifth frame of the test sequences (e.g. in case of Miss America 30 out of 150 frames - Figure 8). This was followed by calculation of the Euclidean distance between the feature tracked manually and automatically. As a result of this, error distance plots were created (examples for the left eye are shown in Figures 8 - 13). Finally, mean error and standard deviation of the error distance plots were calculated. The average difference between manual and automatic tracking was typically less than one pixel. This is comparable to our previous approach using manual pre-fitting of the first *M* frames [8]. This clearly demonstrates the high accuracy of the automatic fitting technique.

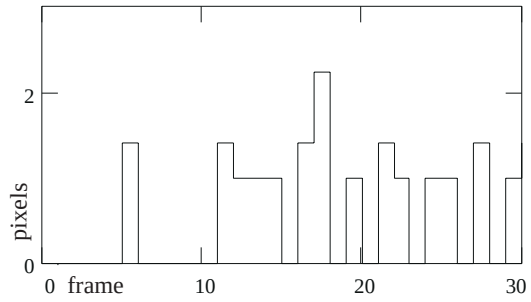

Figure 8. *Miss America* error profile

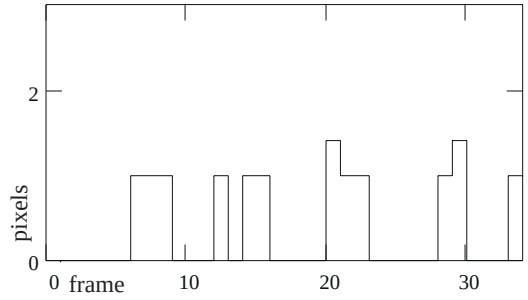

Figure 9. *Claire* error profile

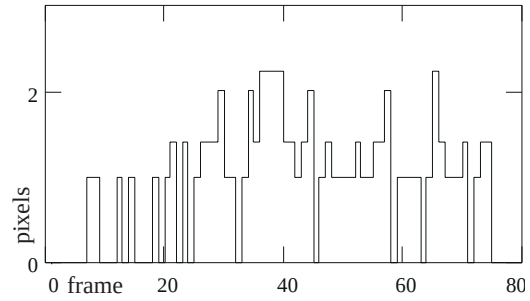

Figure 10. *Car Phone* error profile

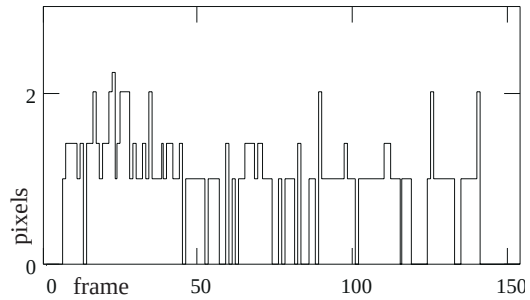

Figure 11. *Grandma* error profile

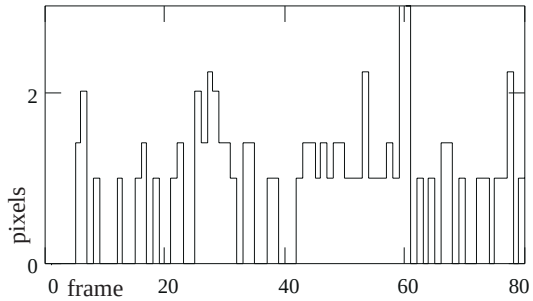

Figure 12. *Salesman* error profile

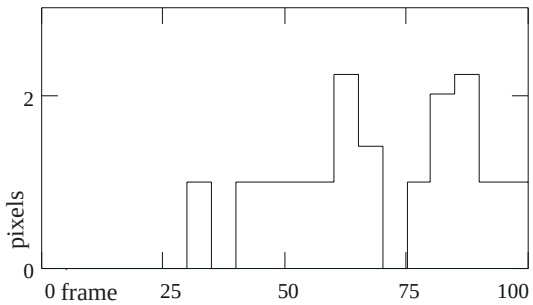

Figure 13. *Trevor* error profile

4. CONCLUSIONS

We have developed a fast and reliable algorithm for fitting and tracking of facial features in 'head-and-shoulders' scenes. The algorithm incorporates significant improvements over previous methods since both fitting and subsequent tracking are performed automatically. This approach uses separate principal component spaces created using sequences of sub-images containing faces and facial features extracted off-line from an independent data-base of images. The algorithm currently processes two frames per second using a machine with the power of a single Pentium processor. Since all the vertices are tracked independently the algorithm can be easily implemented as parallel processes.

In further research we intend to focus on reconstruction of the encoded sequence using texture mapping techniques. Our ultimate goal is to test the applicability of the method for transmission of real-time video at very low data rates.

5. ACKNOWLEDGMENT

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

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] 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.

[8] Antoszczyszyn P.M., Hannah J.M., and Grant, P.M. "Facial features motion analysis for wire-frame tracking in model-based moving image coding", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 21-24 April 1997, vol. IV, pp. 2669-2672.

[9] MIT facial data-base is available via anonymous ftp from the white.media.mit.edu server.

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