AUTOMATIC FRAME FITTING FOR SEMANTIC-BASED MOVING IMAGE CODING USING A FACIAL CODE-BOOK

Paul M. Antoszczyszyn, John M. Hannah and Peter M. Grant
Department of Electrical Engineering, The University of Edinburgh
Edinburgh, EH9 3JL, UK
Tel: +44 131 6505655; fax: +44 131 6506554
e-mail: plma@ee.ed.ac.uk

ABSTRACT

An entirely new method of automatic wire-frame fitting for semantic-based moving image coding is proposed. The algorithm utilises a code-book of facial images. All elements of the facial data-base are pre-processed and manually fitted with the wire frame model. Both pre-processing and manual fitting are a part of the facial images data-base preparation. As such, they are not a part of on-line processing of an unknown image. Only the pre-processed images (monochrome bitmaps) are used in automatic frame fitting. This allows a reduced space requirement for storage of the reference data-base.

1 INTRODUCTION

The issue of extremely low data rate (below 9600 bits/s) moving picture coding has been discussed for over a decade now. Successful implementation of a codec working at low data-rates would allow video communication using standard telephone lines.
It seems unlikely that it will be possible to construct a codec targeting such extremely low data rates, based on waveform coding schemes. The application of model-based coding seems inevitable. We therefore concentrate on knowledge-based (semantic-based) techniques. The reports of Aizawa and Harashima [1] and Forchheimer and Kronander [2] suggest that it is possible to obtain data rates under 10 kbits/s using the concept of a wire-frame fitted to the subject's face. In waveform-based techniques (H. 261, H. 263, MPEG, MPEG II) we deal with spatio-temporal and statistic redundancies. The biggest challenge in the knowledge-based coding approach is the reduction of semantic redundancy. In the semantic-based coding techniques, the actual image of the speaker is sent to the receiver only once (e.g. at initialisation of the videophone link). It is subsequently superimposed onto the wire-frame model using texture mapping. This effectively reduces the semantic redundancy. Because the model is shared by receiver and transmitter, only the parameters of the wire-frame motion need to be sent. This allows reduced data-rate transmission. The two main problems that remain unsolved in semantic-based techniques are scene tracing and automatic wire-frame fitting. This paper suggests a solution for the latter problem.
In most currently proposed techniques, the wire-frame is fitted manually to the subject's face. This solution is unacceptable for obvious reasons. A limited number of approaches to automatic frame fitting have been suggested. The method proposed by Welsh [3] utilises the idea of 'snakes' (active contours). A different approach was presented by Reinders et al [4]. This method is heavily dependent upon the geometry of the face. Seferidis [5] proposed pre-processing using mathematical morphology and actual fitting using chamfer matching. In this paper an entirely new method of automatic wire-frame fitting based on analysis of a facial database is proposed.

2 DESCRIPTION OF THE ALGORITHM

The proposed method is based on analysis of a facial images code-book. In this case the MIT facial data-base was used. Each image in the code-book is pre-processed and pre-fitted with a wire-frame.
The pre-processing is applied in order to simplify the contents of the image. Unlike recognition systems, the automatic wire-frame fitting algorithm is not required to choose the image most similar to the analysed one. The task of the algorithm is to find the image that has the most similar wire-frame fitted to it. Here, the pre-processing consists of three stages: histogram equalisation, isotropic edge detection (using templates as in Table 1)

2	2	2	-2	-2	-2
0	0	0	0	0	0
-2	-2	-2	2	2	2
2	0	-2	-2	0	2
2	0	-2	-2	0	2
2	0	-2	-2	0	2

Table 1: Isotropic templates for edge detector

and thresholding in the middle of the grey-scale range. Only the pre-processed form of the facial images data-base is used in further processing. Because

of that the space required for storage of the reference (monochrome) images is minimal. The pre-fitting is performed manually using a model based on the 'Candide' wire-frame [2].

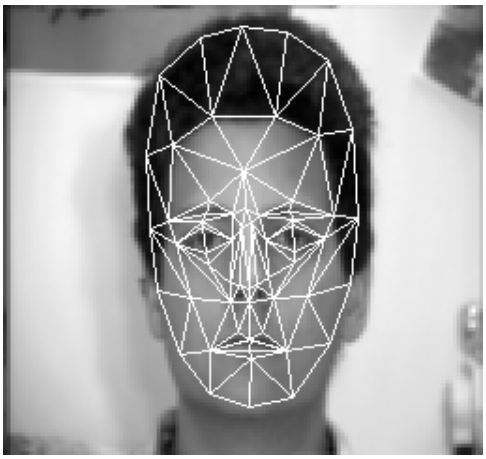

Figure 1: Wire-frame fitted manually

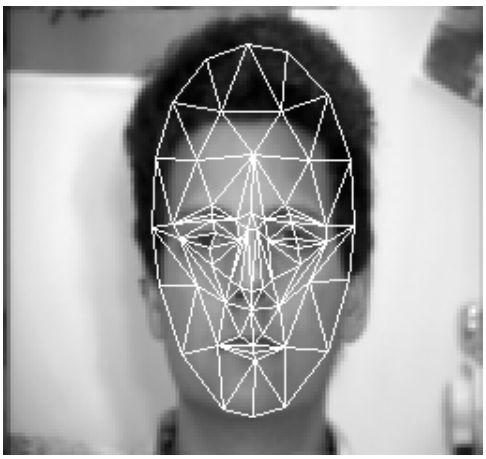

Figure 2: Wire-frame fitted automatically

The quality of manual pre-fitting is judged subjectively by changing the action units [6] (local motion) and three-dimensional co-ordinates of the entire head (global motion). An accurate fit of the wire-frame to facial features is absolutely essential. This is particularly important in the case of lips. The middle of the lips of the subject's face must be exactly mapped onto the middle of the lips of the wire-frame model. The prepared reference data-base is subsequently used in the automatic wire-frame fitting technique.

The unknown (incoming) image is pre-processed using the same algorithm that was applied to the facial images data-base. From that point, the algorithm operates on pre-processed images. The subsequent fitting algorithm consists of two stages. First, the coarse wire-frame fit is performed. The template containing the entire facial area of the code-book image is correlated with the unknown image (the placement and dimensions of the facial area template and facial features templates are derived using the pre-fitted wire-frame co-ordinates). A correlation coefficient R_C is calculated for every possible placement m, n of the template T on the analysed image I using the following equation:

$$R_C(m,n) = \sum_{l=0}^{L} \sum_{k=0}^{K} I(m+k, n+l) * T(k,l)$$

$$0 < m < M - K, \quad 0 < n < N - L$$

where L and K are the dimensions of the template and M and N are the dimensions of the analysed image. The peak in the correlation (maximum value of R_C) estimates the coarse placement of the wire-frame. The correlation process, although performed over the entire image, is relatively fast, since the processed images are by now monochrome bitmaps, and the correlation involves incrementation only.

As soon as the coarse fit is completed the semantic information can be utilised. Here the information about the geometry of the human face and correlation map is combined in the concept of 'floating templates'. The templates of the facial features (eyes, nose and lips) are allowed to float over certain restricted areas of the image. The template of the lips is allowed to drift vertically and horizontally in a window which is twice the area of the lips template itself. The template of the eyes floats in the window which is half of the area of the eyes' template itself. Similar restrictions are applied for the template of the nose. Again, the correlation peak signifies the final placement of the facial feature. The facial features can be chosen from any image in the data-base. It does not need to be the case, that the oval of the face, eyes, nose and lips templates come from the same code-book image. This approach allows a reduced size of data-base and provides increased wire-frame fitting accuracy.

3 EXPERIMENTAL RESULTS

The test images used to verify the proposed algorithm also came from the MIT facial data-base. The results of application of the described algorithm to the 'Joel' image are shown on Figure 2. In order to assess the accuracy of the algorithm, the image was also fitted manually with the same wire-frame (Figure 1). The images presented in Figures 3 to 9 show results of the application of particular action units to both models (manually fitted face on the left, automatically fitted face on the right). The action units and global motion vectors used to test the models are listed in Table 2.

It can be seen, that there is no significant difference between the images presented in pairs. This suggests that the proposed algorithm works reliably (i.e. the quality of automatic wire-frame fit is comparable

Action unit	Level	Figure
Upper lip raiser	+0.5	3
Lip stretcher	-0.5	4
Lip stretcher	+0.7	5
Brow lowerer	+0.3	6
Lip corner depressor	+0.5	7
Outer brow raiser	+0.5	8
Eye closed	+1.0	9
Rotation about ear axis	+0.2	10

Table 2: Action units and global motion vectors used to test the accuracy of wire-frame fitting

to the quality of manual wire-frame fit). It is worthy noticing, that the middle of the lips of the subject was matched to the middle of the lips on the model with very good accuracy. The lips match is crucial for achieving natural-looking texture mapping. The face still looks natural and the quality of manual and automatic wire-frame fitting is comparable. Automatic frame fitting currently takes less than one minute on a 486-40MHz PC computer for a reference data-base of 16 images (128x120 pixels). The code was written entirely in the C++ language. The algorithm would be very easy to implement in hardware. The images used to create the data-base can be obtained via anonymous ftp from: ftp://whitechapel.media.mit.edu/pub/images.

4 CONCLUSIONS AND FUTURE WORK

The results obtained are encouraging and illustrate the success of the technique. Further improvements to the algorithm are possible. Because the accuracy of the wire-frame fit in the lips' area is of crucial importance, new options are under consideration. For example, each template could be weighted to allow more reliable coarse fitting. The biggest weight would be assigned to the template of the lips. Weighting of the templates would result in stronger correlation peaks occurring when the template of the lips of the image from the database passes over the lips' area of the unknown image.

5 ACKNOWLEDGEMENTS

Paul Antoszczyszyn acknowledges the support of the EU TEMPUS Office, Contract No. IMG-94-PL-2325.

References

[1] K. Aizawa and H. Harashima, "Model-based analysis synthesis image coding (mbasic) system for a person's face," *Signal Processing: Image Communication*, vol. 1, pp. 139–152, October 1989.

[2] R. Forchheimer and T.Kronander, "Image coding - from waveforms to animation," *IEEE Transactions in Acoustics, Speech and Signal Processing*, vol. 37, pp. 2008–1023, December 1989.

[3] B. Welsh, "Model-based coding of video images," *Electronics and Comms. Eng. Journal*, vol. 3, pp. 29–38, January 1991.

[4] M. Reinders, P. van Beek, B. Sankur, and J. van der Lubbe, "Facial feature localisation and adapation of a generic face model for model-based coding," *Signal Processing: Image Communication*, vol. 7, pp. 57–74, March 1995.

[5] V. Seferidis, "Facial feature estimation for model-based coding," *Electronics Letters*, vol. 27, pp. 2226–2228, November 1991.

[6] P.Ekman and W. V. Friesen, *Facial action coding system*. Palo Alto, California: Consulting Psychologists Press Inc., 1978.

Figure 3: Upper lip raiser

Figure 4: Lip stretcher

Figure 5: Lip stretcher

Figure 6: Brow lowerer

Figure 7: Lip corner depressor

Figure 8: Outer brow raiser

Figure 9: Eye closed

Figure 10: Rotation about ear axis