IMPROVED TRACKING OF FACIAL FEATURES IN HEAD AND SHOULDER VIDEO SEQUENCES

Somnath Sengupta, John M.Hannah and Peter M.Grant
Department of Electronics and Electrical Engineering,
The University of Edinburgh, King's Buildings, Mayfield Road,
Edinburgh EH9 3JL, Scotland, UK
Contact: Somnath.Sengupta@ee.ed.ac.uk

ABSTRACT

This paper describes work which is being undertaken to develop a low bit-rate switchable video codec for videophone type applications. It describes improvements to the tracking of important facial features, such as the eyes, the nose and the lips in head and shoulder video sequences. Results of investigations into a reference set updating strategy for PCA based tracking are presented. These show that adoption of a suitable updating method can significantly improve the performance of this tracking technique where significant head movement, including rotation, is present. Our proposed technique also uses the relative positions of tracked features to allow detection of frontal or non-frontal head position for use in switching between model based or enhanced feature and waveform based coding.

1 Introduction

Detection and tracking of facial features, such as the eyes the mouth and the lips, is a requirement for low bit-rate model-based video coding of head and shoulders sequences [1]. Other techniques, such as selective enhancement of facial feature regions [2] also depend on accurate facial feature tracking.

Among the many techniques which have been proposed for tracking facial features. is the approach based on Principal Components Analysis (PCA) [3, 4]. This has been shown to give good results with a wide range of common head and shoulders videophone sequences including moderate amounts of speaker's head pan, rotation and zoom [5]. We have therefore adopted this PCA based tracking technique in our ongoing research into low bit-rate video codecs.

In real life, head and shoulders video sequences can be expected to present significant challenges to any feature tracking technique, due to large global movements of the speaker's head and the resultant occlusion of desired features. We have therefore been developing a low bandwidth video codec which will provide enhancement of regions of interest (ROI) [2] when facial features are trackable and use a standard waveform-based approach adopted for the waveform based method [6] (although H.263+ or H.26L could be used). Facial feature enhancement would therefore typically operate when the speaker's face was close to a frontal view and be turned off in non-frontal frames. Frames containing significant rotation of the speaker's head in either the vertical or horizontal planes are classified as non-frontal. Automatic detection of such frames and the ability of the tracking algorithm to maintain the tracking effectively, when the face changes back to frontal after a sequence of non-frontal, frames are both required for a switchable coder.

Automated detection of non-frontal frames, uses the relative coordinates of the eyes and the nose. For a perfectly frontal face, the x-coordinate of the nose should be the the mean of the x-coordinates of the eyes and the eye and nose centres should form an isosceles triangle. Deviation of the nose x-coordinate from the mean x-coordinate of the left eye and the right eye would indicate rotation in the horizontal plane. A significant difference between nose-to-lefteye and nose-to-righteye distances characterises rotation in the vertical plane. The nose, rather than the lips, is used as a reference point for these deviation measures, since the lips may change shape significantly, whereas the nose is a relatively rigid feature, making its accurate localisation easier.

Where possible, feature tracking should be continued during a sequence of non-frontal frames. When the face turns to a semi-profile or profile position, at least one of the eyes should still be visible and hence trackable, whereas the other eye may be partly occluded and may not be tracked reliably. Thus, when a sequence of non-frontal frames is detected, a further facial geometry consistency check on the tracked eye position is adopted. When this is satisfied and the frontal measures mentioned above indicate a frontal face frame, normal tracking can be restored.

Our use of a PCA based technique for feature tracking in video sequences containing significant rotational motion of the head has led us to investigate ways of improving the performance of this approach. Previous work [5] has used a fixed reference set of eigenfeatures because this was found to give the best performance. The issues associated with methods of updating the reference eigenspace are similar to those described by Peacock *et al* [7] on reference block updating when tracking with block matching. This paper presents some results of our investigation into improvements in PCA based facial feature tracking. It proposes techniques which can lead to significant performance enhancement through up-

dating the reference eigenspace in an appropriate manner.

2 The Tracking Algorithm

The tracking of facial features, employed in this approach is based on Principal Component Analysis (PCA). As a first step, the eigenvectors of the covariance matrix S of the sequence X of M, N-dimensional input column vectors of the input matrix : $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M], \mathbf{x}_j = [\mathbf{x}_{ji}], i = 1, ...N, j = 1, ..., M$ must be found. The covariance matrix can be obtained from the following relationship :

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

where the columns of matrix $\mathbf{Y}$ are vectors $\mathbf{y}_j$ that differ from the input vectors $\mathbf{x}_j$ by the expected value $\mathbf{m}_x$ of the sequence $\mathbf{X}$:

$$\mathbf{Y} = [\mathbf{x}_1 - \mathbf{m}_x \mathbf{x}_2 - \mathbf{m}_x ... \mathbf{x}_M - \mathbf{m}_x] \tag{2}$$

$$\mathbf{m}_x = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}_i \tag{3}$$

The eigenvectors $\mathbf{u}_i, i = 1, ..., N$ of the covariance matrix $\mathbf{S}$ can be calculated from the following relationship :

$$\mathbf{S}\mathbf{u}_i = l_i \mathbf{u}_i \tag{4}$$

where, l_i is the eigenvalue of the ith eigenvector, $i = 1, ..., N$.

The input matrix X is formed from a reference set of sub-images containing important facial features. There is a separate reference set for each tracked facial feature such as the left eye, right eye, lips and nose. In our tracking approach, each facial feature, namely the left eye, the right eye, the nose and the lips are tracked separately. Each reference set contains M sub-images. The sizes (N) of each of these sub-images is dependent on which facial feature is to be tracked. Tracking proceeds by extracting a set of sub-images, within a reasonable search range, from the current frame and selecting the sub-image from the extracted set that is most similar to all the images from the reference set, using the measure proposed in [5]. This finds the image that is most similar to all the images from the reference set in principal component space.

In the previous work, the reference set is created from the first M (normally 16) frames of the video sequence to form the input matrix $\mathbf{X}$. This works satisfactorily as long the rest of the video sequence contains frames having broadly similar facial features. However, this may not be always the case where significant facial pose deviations and varying feature conditions are involved. Therefore, we have investigated a reference set updating policy, so that the matrix $\mathbf{X}$ and consequently the matrix $\mathbf{U}$ formed using eigenvectors $\mathbf{u}_i$, $i = 1, ..., N$ is updated before feature tracking in each frame. This obviously increases the overall processing required but can give improved performance as will be shown below.

The detection of non-frontal face frames is based on two simple frontal deviation measures. The first measure D_1 is defined as

$$D_1 = \frac{x_N - \frac{x_{LE}+x_{RE}}{2}}{x_{LE} - x_{RE}} \tag{5}$$

where $(x_{LE}, y_{LE}), (x_{RE}, y_{RE})$ and (x_N, y_N) are the coordinates of the left eye, the right eye and the nose respectively. This calculates the offset between the x-coordinate of the nose and the mean x-coordinates of the two eyes, normalised by the separation between the two eyes. This measure is not only sensitive to head rotation in the horizontal plane but also sensitive to rotation in the vertical plane. A second measure is also used which determines the differences between the nose to eye distances and is given by

$$D_2 = \frac{d((x_{LE}, y_{LE}), (x_N, y_N)) - d((x_{RE}, y_{RE}), (x_N, y_N))}{d((x_{LE}, y_{LE}), (x_{RE}, y_{RE}))} \tag{6}$$

where $d(.,.)$ is the Euclidean distance between the pair of points.

A frame is classified as non-frontal if $\mid D_1 \mid \geq \theta_1$ and $\mid D_2 \mid \geq \theta_2$, where θ_1 and θ_2 are thresholds. Any updating of the eigenspaces of feature sub-images is stopped during non-frontal frames. Beginning with the next frame, the tracker globally searches for the desired facial features in this frame. Since the feature positions detected through such global search may not be reliable, as one or more of the features may be partially or fully occluded because of head rotation, further facial geometry based consistence checks are introduced:

1. The left eye should be located to the left of the right eye. This condition is important because sometimes occlusion of one of the eyes may cause a false match with the other eye.

2. The left eye and the right eye should be to the left and the right respectively of the detected nose position. This condition is essential to achieve re-tracking the frontal frames.

3. The vertical directional separation between the left eye and the right eye should be lower than a reasonable threshold.

4. The eyes shouldn't be too close nor too distant, using realistic thresholds.

5. The lips should be located below the nose and the eyes should be located above the nose.

Obviously such global searching for features within a frame is computationally intensive. Thus when the frontal conditions for D_1 and D_2 and the consistency checks are satisfied the tracker reports a successful frontal track and subsequent frames are tracked with only local searching, as before.

3 Results and Discussion

Our tracking algorithm was tested with the commonly used "Foreman" sequence in QCIF (image size : 176x144 pixels) format. As illustrated in fig. 1, the foreman sequence exhibits many varieties of facial poses, such as fully frontal, tilt up , left turn, lip occlusion, right turn, tilt down, demonstrated in (a) to (f). It is thus a challenging sequence on which to test our algorithm. in addition, we had manually recorded the feature positions and labelled the frame as frontal or non-frontal for each and every frame for the first 280 frames in the sequence, which enabled us to obtain quantitative results.

We considered and experimented with a wide range of different updating strategies for the reference eigenspace, for our PCA based tracking algorithm:

1. **Fixed Eigenspace with features from the first 16 frames**. This was the method found to be most successful previously [5]. It was used along with the frontal deviation measures and consistency checks described above.

2. **Fixed Eigenspace with diverse feature templates**. This used the same algorithm, but the feature templates were carefully selected manually from 16 separate frames in the sequence, so that they represented a range of typical views. Since a selection like this cannot be easily automated in practice, this experiment was conducted to ascertain how our algorithm would perform with a diverse feature set. It therefore provided a 'benchmark' with which to compare other strategies.

3. **Best Match Replacement**. Here, the detected feature template from the current image is compared with the images in the reference set that is used to form the eigenspace. This feature template is used to replace the existing one at the best matching position in the mean-squared error sense. This updating strategy attempts to ensure that the reference set does not simply consist of very similar feature templates.

4. **Best Match Replacement in every 5th frame**. This is the same as above but the update process only take place in every 5th frame. This also aims to maintain a fairly diverse reference set.

5. **Oldest Replacement**. The detected feature template from the search image replaces the oldest out of the sixteen templates in the feature sub-image set. This is a very simple strategy which attempts to adapt the reference set to feature changes.

6. **Oldest Replacement in every 5th frame**. Same as above but replacement only occurs in every 5th frame. This attempts to follow changes but avoid too similar a reference set.

7. **Fixed + Oldest**. This updating strategy is a hybrid of the fixed and updated eigenspaces. The first eight templates are kept fixed and the remaining eight are updated in accordance with the oldest replacement. This therefore achieves a combination of original as well as recent feature templates.

8. **Fixed + Best**. This is similar to the above, but here the last eight templates are updated in accordance with the best replacement strategy.

Our tracking algorithm with these different updating strategies was run on the foreman sequence and the results are tabulated in table- 1. This shows the mean as well as the standard deviations of the lefteye, the righteye, the lips and the nose are shown in terms of pixel distances from the manually tracked positions for the various methods. These allow the relative tracking accuracy of the various updating strategies to be compared with each other. It should be noted that the manual tracking accuracy is not better than +/- 1 pixel. The table also shows, in the penultimate columns the percentage of frontal frames (manually determined as 190) which were successfully tracked. The final column gives the percentage of total frames (264) which were tracked. The first 16 frames were excluded from these statistics.

As might be expected, the fixed eigenspace with a manually selected diverse template set performs best in terms of average feature error and percentage of tracked frontal frames. The updating strategy that most closely approaches the performance of this 'idealised' technique is the "First + Oldest" method. It appears that dividing the reference eigenspace into two halves, one of which is updated and one which is not, provides a good compromise. It seems that the most recently acquired eight templates help in tracking frames closer to the recent ones, whereas preserving the first eight improves the ability to track frontal frames which are significantly different from the recent ones. Interestingly, the "Best Match Replacement" strategy achieves similar tracking accuracy to the "Fixed + Oldest" strategy while exhibiting the best performance in terms of the percentage of total frames tracked. This probably results from its ability to track some non-frontal frames, although it is slightly less successful at frontal tracking.

It should be noted that in all the updating strategies, the updating of feature templates was stopped during 'non-frontal' tracking, so that erroneously tracked features do not affect the reference eigenspace and the subsequent tracking results.

The new algorithm was implemented and tested on SUN Ultra-10 Workstation with Solaris-8 Operating System at 700 MHz. Tracking of each frame took 0.7 second on an average and it is mostly contributed by the exhaustive search for best matching feature that we performed within a window range of +/- 6 pixels. For a real-time implementation, special hrdware architecture must be designed.

4 Conclusions

This paper has described improved tracking of facial features for head-and-shoulder video sequences containing non-frontal scenes. Our tracking algorithm recovers the tracking

Eigenspace	Lefteye error		Righteye error		Lip error		Nose error		% Frontal tracked	% Total tracked
	Mean	σ	Mean	σ	Mean	σ	Mean	σ		
Fixed (initial)	1.44	1.37	1.61	1.82	3.49	3.73	1.49	1.57	89.47	72.86
Fixed (diverse)	1.44	1.16	1.57	1.78	1.28	0.99	1.02	1.56	96.84	81.82
Best Match Replacement	1.91	1.37	2.04	1.96	1.80	1.82	1.65	2.58	90.00	83.27
Best Match every 5th	1.20	1.07	1.60	1.77	3.87	3.32	1.45	1.63	84.74	69.32
Oldest Replacement	4.43	4.30	6.30	6.10	1.42	1.33	2.18	0.92	75.79	67.42
Oldest every 5th	1.25	1.00	1.69	2.08	4.46	3.86	2.40	2.33	85.26	71.59
Fixed + Oldest	1.82	1.49	1.98	1.85	1.76	1.41	1.55	1.52	95.26	80.68
Fixed + Best	1.57	1.11	2.31	2.28	1.82	1.83	1.49	2.41	90.52	78.79

Table 1: Performance comparison of different updating policies on Foreman sequence upto frame no.280

(a) (b) (c)

(d) (e) (f)

Figure 1: Examples of facial poses of "Foreman" (a) frame no.16, (b) frame no.68 (c) frame no. 188 (d) frame no.255 (e) frame no.265 (f) frame no. 279

efficiently after sequences of non-frontal frames. Various updating strategies for the reference eigenspace in PCA-based tracking have been considered and compared. The results show that a suitable updating strategy can give significant performance improvements and good robustness.

Other possible techniques for improving tracking might have included increasing the size of the reference set or using multiple eigen spaces. However, these would have significantly increased the computational complexity. We believe that the approach we have adopted offers good results within modest computing requirements.

Our improved tracking techniques are being adopted for a switchable codec which can efficiently switch between standard waveform based and model/foveation based coding for optimal low-bit-rate coding performance.

The tracking algorithm is implemented in software and it is far from being real time. Designing suitable architecture for this tracker is a topic for further research.

5 Acknowledgements

This work is supported by EPSRC grant number GR/M57255.

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 Communication*, vol. 1, no. 2, pp. 139–152, October 1989.

[2] S.Sengupta, J.M.Hannah, and P.M.Grant, "Enhancing selected facial features in very low bit-rate video sequences," in *IEEE Int. Conf. on Image Processing (ICIP), Thessaloniki, Greece*, 2001, vol. I, pp. 485–488.

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

[4] A.Pentland, B.Moghaddam, and T.Starner, "View-based and modular eigenspaces for face recognition," in *Proc.IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington*, June 1994.

[5] P.M.Antoszczyszyn, J.M.Hannah, and P.M.Grant, "A new approach to wire-frame tracking for semantic model-based moving image coding," *Signal Processing : Image Communication*, vol. 15, pp. 567–580, 2000.

[6] S.Sengupta, J.M.Hannah, and P.M.Grant, "Improving the quality of very low bit-rate video by selective quantization of facial features," in *Proceedings of International*

Workshop on Very Low Bitrate Video Coding, 2001, pp. 62–65.

[7] A.M. Peacock, S. Matsunaga, D. Renshaw, J.M. Hannah, and A.F Murray, "Reference block updating when tracking with the block matching algorithm," *Electronics Letters*, vol. 36, no. 4, pp. 309–310, 2000.