To appear in: Proc. 1st Int. Conf. on Audio- and Video-based Biometric Person Authentication, Lecture Notes in Computer Science, Springer Verlag, 1997

Tracking Facial Feature Points with Gabor Wavelets and Shape Models

Stephen J. McKenna[1], Shaogang Gong[1], Rolf P. Würtz[2], Jonathan Tanner[1], Daniel Banin[1]

[1] Machine Vision Laboratory, Department of Computer Science, Queen Mary and Westfield College, Mile End Road, London. E-mail: stephen@dcs.qmw.ac.uk
[2] Institute for Mathematics and Computing Science, University of Groningen, P.O. Box 800, NL-9700 AV Groningen, The Netherlands.

Abstract. A feature-based approach to tracking rigid and non-rigid facial motion is described. Feature points are characterised using Gabor wavelets and can be individually tracked by phase-based displacement estimation. In order to achieve robust tracking a flexible shape model is used to impose global constraints upon the local feature points and to constrain the tracker. While there are many applications in facial analysis, the approach can be used for tracking other textured objects.

1 Introduction

The ability to track facial motion is useful in applications such as face-based biometric person authentication, expression analysis, animation and teleconferencing. In particular, person identification based upon image sequences can make use of temporal information and is potentially more reliable than systems which base their outcome upon a set of 'snap-shot' images.

This paper describes an approach to facial motion estimation which tracks facial feature points. It combines local image measurements in the form of Gabor wavelets with a global shape model realised as a Point Distribution Model (PDM). The approach is generic and can easily be applied to textured objects other than faces. This is because it does not employ any detailed or overly specific 3D head model or facial muscle model. The models are instead built from example images.

Gabor feature jets have previously been used for static face analysis with elastic graph matching [8] and coarse-to-fine correspondence matching [14]. Graphs with different topologies were needed in order to handle different head poses [7]. Gabor feature jets have recently been used to track facial feature points on faces rotating in depth [10]. However, each point was treated independently with no global shape model to constrain the tracking. The method was, therefore, susceptible to the inevitable tracking errors which occur due to aperture problems, noise and occlusions.

Research carried out at the University of Manchester has involved extensive use of PDMs to model the possible shape variations of faces. The image measurements used typically correspond to grey-level profile or gradient-based features

extracted along normals to a contour [9]. PDMs have also been used for tracking contours [2].

The method proposed here combines the advantages of Gabor wavelets for feature characterisation and displacement estimation with a PDM in order to achieve robust tracking. In the next section the use of a 'jet' of Gabor wavelets to track a feature point is described. In Section 3, the PDM which models the possible shape variations of faces is described. Section 4 outlines the integration of the Gabor wavelet-based tracking with the PDM. Some preliminary results are presented in Section 5 and concluding remarks are made in Section 6.

2 Feature Tracking

For robust tracking of single points the choice of local features is crucial. Grey values, the most local features possible, suffer from huge ambiguities as well as sensitivity to camera noise and slight changes in illumination. More robust features can be obtained by a local combination of several pixels, or filtering. The combination of several filter responses into a feature vector yields even better results. In particular, for correspondences between textured objects such as human faces, *Gabor filters* have been shown to be very effective [14]. They have the following form.

$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{\mathbf{k}^2}{\sigma^2} \exp\left(-\frac{\mathbf{k}^2\mathbf{x}^2}{2\sigma^2}\right) \left[\exp\left(-i\mathbf{k}\mathbf{x}\right) - \exp\left(\sigma^{-2}\right)\right]$$

Feature vectors, or 'jets', are built from the local responses of these filters at an image location $\mathbf{x}$. The components of a jet correspond to different *centre frequencies* $\mathbf{k}_i$. (The two-dimensional vector $\mathbf{k}_i$ governs spatial frequency and orientation of the filter). The parameter σ controls the width of the Gaussian window relative to the wavelength corresponding to $\mathbf{k}_i$. The fact that this ratio is constant turns the convolution with this set of filters into a (nonorthogonal) wavelet transform. The complex formulation of the Gabor filters has the advantage that it is natural to split up the responses into amplitude (or modulus) and phase rather than real (even) and imaginary (odd) parts. Both these amplitudes and phases can be used as *Gabor features* in establishing correspondences with different matching heuristics. The amplitudes provide smooth similarity "landscapes" with few local minima, but the localization of the minima is poor. Thus, good matches of two corresponding Gabor features can be found by stochastic search procedures [8] or template matching in low resolution [14, 15]. The similarity landscapes for phases are very ragged but relatively easy to predict: in the absence of very small amplitudes they behave in a roughly linear way, i.e. the phase in the direction of the centre frequency approximately rotates with that frequency [6]. Thus, a phase difference in a component of the feature vector can be translated into a local displacement in the corresponding direction. These feature displacement estimates for each direction were combined into one displacement using a least squared error criterion.

A simple measure of the saliency, ξ, of a feature point was defined in terms of the amplitudes, a_j, of the Gabor filter responses at that point [13]:

$$\xi = \sqrt{\sum_j a_j^2} \tag{1}$$

Displacements were only estimated for salient feature points i.e. those with ξ larger than a predefined threshold. An initial estimate of a feature's displacement was based upon the phase responses, ϕ_j, of the Gabor wavelets at the lowest spatial frequency level used in the jets. The feature's position was then updated and a further displacement from this position was estimated using Gabor wavelets at the next highest spatial frequency. Further iterations at even higher frequencies achieved increasingly precise estimates. In practice, only low frequency Gabor responses were needed because the phase information yielded sub-pixel accuracy. The estimated feature displacement, $\mathbf{d}$, was assigned a confidence measure proportional to the similarity $S(J, J')$ between the jets before (J') and after (J) displacement, where [13]:

$$S(J, J') = \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j - \mathbf{dk}_j)}{\sqrt{\sum_j a_j^2 \sum_j a_j'^2}} \tag{2}$$

Features on the face which typically had high saliency were selected for tracking. In the current implementation, feature positions were initialised by hand in the first frame of each sequence. However, given a frontal view of the head and a rough localization of the face, a coarse-to-fine matching process could have been performed to locate facial features and to allocate nodes at these features [14].

A disadvantage with this kind of purely feature-based tracking is that errors in the displacement estimates accumulate and nodes loose lock on their corresponding features. Aperture problems mean that many features will inevitably drift. Such displacement errors need to be compensated using contextual information from correctly matched feature points elsewhere on the face.

3 Shape Model

Point Distribution Models (PDMs) provide a method for representing flexible objects by means of a set of feature points describing a deformable shape [4]. The aim of a PDM is to capture the flexibility of objects in a compact manner. This section briefly describes the PDMs used.

An object's shape is represented as a $2n$-dimensional vector of n image coordinates, $\mathbf{s} = (x_0, y_0, x_1, y_1, \ldots, x_{n-1}, y_{n-1})$. In the case of contour tracking, a B-spline curve with these points as control points is usually used to delineate the contour [2]. Here, however, the PDM is not used to perform contour tracking, rather, $\mathbf{s}$ corresponds directly to the locations of Gabor feature jets which need not lie on any meaningful contours.

PDMs were built from a set of training shapes captured by labelling an ordered set of feature points on images from video sequences of faces undergoing changes in pose and expression. Translation, scaling and rotation in the image-plane were not modelled directly by the PDM. These transformations were instead normalised using a separate object alignment algorithm [4]. Let R_{kl} be the distance between two feature points (x_k, y_k) and (x_l, y_l) in the shape model and let $V_{R_{kl}}$ be the variance of this distance over the training set. A weight w_k for each point k was assigned using

$$w_k = (\sum_{l=0}^{n-1} V_{R_{kl}})^{-1}$$

Feature points which were stable relative to other feature points were assigned large weights and their alignment was thus given greater priority. Alignment was achieved by rotating, translating and scaling shapes in order to minimise the weighted sum-of-squares distance between them. The iterative alignment of the training set proceeded as follows:

1. All training shapes were rotated, translated and scaled into alignment with $\mathbf{s}_0$, the first shape in the set
2. The mean of these transformed shapes was then aligned with the first shape in the set
3. All training shapes were rotated, translated and scaled into alignment with the mean shape
4. Steps 2 and 3 were repeated until convergence

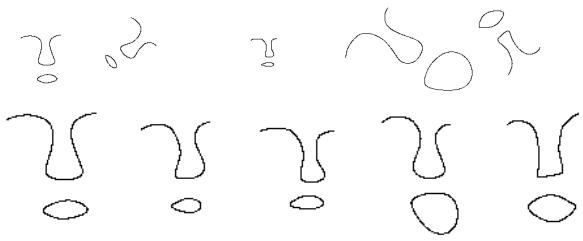

Fig. 1. An example training set before alignment (top row) and after alignment (bottom row). The feature points have been used to draw B-spline curves solely for illustration purposes.

Principal components analysis of the aligned training set then allowed shape vectors to be approximated in terms of the mean training shape, $\bar{\mathbf{s}}$, and the eigenvector matrix, L, which captured the modes of variation in the training set:

$$\mathbf{s}' = L\mathbf{b} + \bar{\mathbf{s}} \tag{3}$$

The shapes, $\mathbf{s}'$, generated by this model were constrained to lie within a hyper-ellipsoid by restricting the permissible values of $\mathbf{b}$ such that each element b_i was in the range $[-2.5\sqrt{\lambda_i}, 2.5\sqrt{\lambda_i}]$, where λ_i was the i^{th} eigenvalue [4].

A generic model can be obtained by training on a large set of images of different people. Alternatively, a person-specific model is simpler to train and allows more robust tracking for that person.

4 Tracking Facial Feature Points with Shape Models

Tracking was initialised by manually positioning the feature points of the face shape model onto the first frame of each sequence. In each subsequent frame, t, new feature positions $\mathbf{x}(t)$ were assigned as follows. An estimate $\tilde{\mathbf{x}}(t)$ was obtained using Gabor wavelet-based displacement estimation from $\mathbf{x}(t-1)$ as described in Section 2. This estimate was then aligned with the PDM shape model and projected onto the eigenvectors, L, to yield a vector $\mathbf{b}$. The elements of $\mathbf{b}$ were then constrained as described in Section 3 and Equation (3) was used to reconstruct an approximation to the shape. This reconstructed shape was then re-aligned with the image to give the new feature positions $\mathbf{x}(t)$. Feature points with low confidence Gabor displacement estimates were disregarded during alignment and reconstruction using the PDM [1].

The PDM had the effect of constraining the tracked shape to lie within a hyperellipsoid thus helping to ensure that it was a valid face shape. In addition, the PDM 'regularised' the shape at each frame enabling erroneous displacement estimations to be corrected.

Non-rigid distortions of the face and moderate head rotations in depth were modelled using a single PDM. Since the PDM is a linear model, it can only give an approximation to the actual shape deformations caused by this combined rigid and non-rigid motion. However, the PDM was found to cope adequately.

The convolutions required to implement the Gabor wavelet feature jets represented the most computationally expensive component of the tracker. An approximation to the Gabor wavelet transform was implemented using two convolution devices on a Datacube MaxVideo 250 pipeline machine. Thus, face tracking can be performed in real-time.

5 Tracking Results

A 'generic' PDM was trained using images of 6 different people. Some example training images are shown in Figure 2. Thirty-nine facial feature points were used and these were positioned around the eyes and the mouth. The first 9 eigenvectors accounted for 95% of the shape variation and only these were used in the PDM.

Figure 3 shows 6 frames from a test sequence of a person not included in the training set. The sequence was 250 frames long and consisted of a face nodding up and down and then turning from left to right. The eyes and mouth were

Fig. 2. Some of the training images.

Fig. 3. Six frames at 40 frame intervals from a tracked face sequence.

successfully tracked through changes in head pose. The final frame shows the tracker dealing with the mouth opening. After 250 frames, the tracker had not lost lock although some features around the right eye were slightly misplaced.

6 Discussions and Future Work

A method for tracking facial motion was described. Gabor wavelet jets were used to estimate feature displacements and a PDM was used to constrain tracking by imposing global knowledge of probable shape variations. Special convolution hardware was used to obtain a real-time implementation for the Gabor wavelets.

It is of course possible to obtain more domain-specific tracking by incorporating specialised models of facial muscle deformation or 3D head shape (e.g. [5]). However, the method presented here used only 2D shape information derived directly from training images. It was, therefore, more generic and applicable to other textured objects. Given suitable training data, tracking was robust over

long sequences. Accumulation of displacement errors was small after 250 frames.

A limitation of the PDMs used is that they are linear making them inappropriate for modelling non-linear effects such as bending or rotation of shape sub-components about one another [4]. Linear PDMs are not sufficiently specific to allow robust tracking of faces undergoing very large deformations and rotations in depth. However, performance was acceptable on the sequences used here. More extensive evaluation with representative face sequences and the use of other facial features is required. Non-linear PDMs have been suggested [11, 12].

In addition to shape information, jet responses at each feature point could be extracted during training. These responses could then be used during tracking to prevent features from drifting. Several different PDMs based on different feature subsets might be used to span the viewing sphere. This would allow self-occlusion under large rotations in depth to be handled. Kalman filtering could be used to track the feature positions, thereby increasing robustness and allowing faster visual motion to be handled. Temporal aspects of deformation might be more directly incorporated using spatio-temporal models (see e.g. [3]).

References

1. D. Banin. Tracking faces with flexible shape models. Master's thesis, Queen Mary and Westfield College, London, September 1996.
2. A. M. Baumberg and D. C. Hogg. An efficient method for contour tracking using active shape models. Technical report, School of Computer Studies, University of Leeds, April 1994.
3. A. M. Baumberg and D. C. Hogg. Learning spatio-temporal models from training examples. Technical report, School of Computer Studies, University of Leeds, March 1995.
4. T. F. Cootes, C. J. Taylor, Cooper D. H., and Graham J. Training models of shape from sets of examples. In *BMVC*, pages 9–18, 1992.
5. D. DeCarlo and D. Metaxas. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *CVPR*, pages 231–238, 1996.
6. David J. Fleet. *Measurement of Image Velocity.* Kluwer Academic Publishers, 1992.
7. N. Kruger, M. Potzsch, and C. von der Malsburg. Determination of face position and pose with a learned representation based on labeled graphs. Technical Report 96-03, Institut fur Neuroinformatik, Ruhr-Universitat Bochum, January 1996.
8. Martin Lades, Jan C. Vorbrüggen, Joachim Buhmann, Jörg Lange, Christoph von der Malsburg, Rolf P. Würtz, and Wolfgang Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.
9. A. Lanitis, C. J. Taylor, T. F. Cootes, and T. Ahmed. Automatic interpretation of human faces and hand gestures using flexible models. In *Int. Workshop Automatic Face and Gesture Recognition*, Zurich, 1995.
10. T. Maurer and C. von der Malsburg. Tracking and learning graphs on image sequences of faces. In *Proc. Int. Conf. on Artificial Neural Networks*, Bochum, 1996.

11. P. D. Sozou, T. F. Cootes, Taylor C. J., and E. C. Di Mauro. A non-linear generalisation of PDMs using polynomial regression. In *5th BMVC*, pages 397–406, York, 1994.
12. P. D. Sozou, T. F. Cootes, Taylor C. J., and E. C. Di Mauro. Non-linear point distribution modelling using a multi-layer perceptron. In *6th BMVC*, pages 107–116, Birmingham, 1995.
13. L. Wiskott. *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis.* PhD thesis, Ruhr-Universitat Bochum, July 1995.
14. R. P. Würtz. *Multilayer Dynamic Link Networks for Establishing Image Point Correspondences and Visual Object Recognition.* Verlag Harri Deutsch, 1995.
15. Rolf P. Würtz. Object recognition robust under translations, deformations and changes in background. *IEEE PAMI*, 1996. Submitted.

This article was processed using the LaTeX macro package with LLNCS style