LEARNING A FACE MODEL FOR TRACKING AND RECOGNITION

Zakaria Ajmal[1] *Jean-Yves Bouguet*[2] *Russell M. Mersereau*[1]

[1]School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
zakaria, rmm@ece.gatech.edu

[2]Intel Corporation – SC12-303
2200 Mission College Blvd
Santa Clara, CA 95054, USA
jean-yves.bouguet@intel.com

ABSTRACT

This paper describes a system for learning a face model that is used for 3D tracking of the human face. The face is modeled as a linear combination of shape basis vectors and action vectors. Shape space models the difference in face shape of different people while action space models the facial expressions. First, real stereo tracking data is used to learn the space of these shape and action vectors using Principal Component Analysis. Then this low-complexity model is used to simultaneously track shape, pose and expression from a monocular image sequence. The main contribution of this paper is in learning shape and action deformation models simultaneously from real data. Results of monocular model-based tracking for subjects not included in the training set show that the model derived from data is robust and generalizes well.

1. INTRODUCTION

Face modelling and tracking are important components in systems for human-computer interaction, video conferencing and model-based video compression. Many researchers have worked on these problems over the last ten years [1, 2, 3, 4, 5].

This paper describes a system that learns the face model and then tracks the face using the learned model. The face model incorporates pose, shape and action deformations. The shape deformations model the differences in shape between different people, while the action deformations model the facial expressions independent of the subject. In the training phase, the model is learned from tracked feature points (eyes, nose, mouth) in the stereo image sequences using Principal Component Analysis (PCA). The model is then used to track the face in a monocular sequence using an optical flow based tracking algorithm.

Most researchers doing model-based face tracking have used hand-crafted models. DeCarlo and Metaxas [1, 6] used optical flow to track pose and action deformations. Their deformable model consists of several articulated points and has to be designed by hand prior to tracking. Eisert and Girod [7, 8] tracked the human face using a similar model-based approach. Ahlberg [5] used a hand-crafted model CANDIDE [9].

Principal Component Analysis is a powerful method for estimating a low-dimensional representation of data from a higher-dimensional space. Bregler et al. [10] and Blanz and Vetter [11] used PCA to approximate non-rigid shapes as linear combinations of sets of rigid basis shapes.

Cootes and Taylor [4] learn an active shape model from training data using PCA. They do not, however, distinguish between shape and action deformations. Also, their tracking algorithm does not use optical flow.

In previous work [12], action deformation was modelled using Principal Component Analysis. A different model was computed for each subject. The models were then used to track that same subject in a monocular sequence.

The rest of the paper is organized as follows. First, we describe the details of our system in Section 2. Then, the experimental results which show the effectiveness and robustness of our approach are given in Section 3. Finally, we discuss possible extensions for future work in Section 4.

2. DESCRIPTION OF THE SYSTEM

Figure 1 shows the main components of the system. In the first stage, the 3D model of the face is learned from stereo image sequences. A low-complexity mesh is initialized on the face and then tracked through the sequence using optical flow techniques. The sequence of meshes for different subjects are aligned and used to compute a compact deformable model of the face using Principal Component Analysis.

In the second stage, the learned face model is used for monocular tracking. The subject is free to change his pose and facial expression. The system uses a model-based optical flow technique to track the face of the subject.

Using real data to construct the model allows us to capture the shape and action deformation spaces with a small, optimum number of basis vectors. Also, the generalization properties of the model can be confirmed by checking the fit of the model and the tracking performance on test subjects not included in the training set.

2.1. Deformable Face Model

The face is modelled by a collection of $N = 19$ points. The face reference frame is attached to the head. Therefore, at frame n, $\mathbf{X}_c^i(n)$, the coordinates of the ith point on the face mesh in the camera reference frame, is related to $\mathbf{X}^i(n)$, the coordinates in the face reference frame, by a rigid transformation.

$$\mathbf{X}_c^i(n) = \mathbf{R}(n)\mathbf{X}^i(n) + \mathbf{t}(n) \qquad \text{for } i = 1, \ldots, N \qquad (1)$$

where $\mathbf{R}(n)$ and $\mathbf{t}(n)$ are the rotation matrix and translation vector for frame n. $\mathbf{R}(n)$ is uniquely parameterized by a rotation vector $\omega(n)$. The rotation vector and the rotation matrix are related by the Rodrigues formula [13].

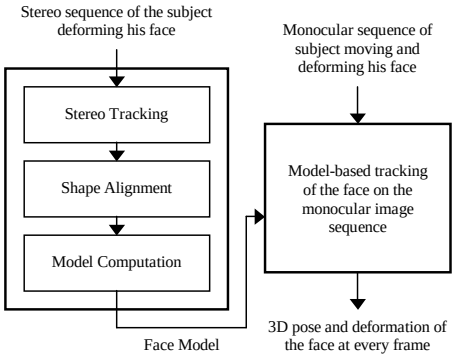


Fig. 1. Main components of the algorithm.

Let $\mathbf{x}^i(n)$ be the 2-D image coordinates of the point $\mathbf{X}^i(n)$. The projection of the 3-D point onto the image is a function of the camera intrinsic and extrinsic parameters.

$$\mathbf{x}^i(n) = \pi_c(\mathbf{X}^i(n), \omega(n), \mathbf{t}(n)) \tag{2}$$

The face is modelled as the sum of a mean shape $\overline{\mathbf{X}}^i$ and a linear combination of shape and action basis vectors.

$$\mathbf{X}^i(n) = \overline{\mathbf{X}}^i + \mathbf{S}^i\sigma(n) + \mathbf{A}^i\alpha(n) \tag{3}$$

where $\mathbf{S}^i$ and $\mathbf{A}^i$ are matrices of size $3 \times p$ and $3 \times q$ respectively (p and q are the respective dimensions of the shape and action spaces.) The columns of $\mathbf{S}^i$ and $\mathbf{A}^i$ are the shape and action basis vectors respectively, and $\sigma(n)$ and $\alpha(n)$ are the shape and action parameter vectors for frame n.

During tracking, the pose parameters ω and $\mathbf{t}$, and the deformation parameters σ and α are estimated from the image sequence under the model constraints of equations 1, 2 and 3. In the training phase, the mean shape $\overline{\mathbf{X}}^i$ and the shape and action basis vectors $\mathbf{S}$ and $\mathbf{A}$ have to be learned from data.

2.2. Stereo Tracking

To train the model, a calibrated stereo camera is used to track the face of a subject in a sequence in which the subject makes a variety of facial expressions without changing his pose.

A set of $N = 19$ points located on the eyes (2), nose (3), mouth (8) and eyebrows (6) are initialized on the first pair of images (See Figure 2). These points are then tracked using 3-D optical flow techniques [14, 15] constrained by space-energy functions as described in [12]. The result of this procedure is the 3-D trajectory of each point $\mathbf{X}^i_c(n)$ through the entire sequence.

2.3. Shape Alignment

The tracked 3-D meshes for every sequence and subject have to be aligned to each other. The most popular method of aligning shapes is Procrustes analysis [16]. This method applies a rigid transformation (rotation and translation) to each mesh to minimize the sum of distances of each shape to the mean. We use Procrustes analysis to first align the meshes for each subject separately to the mesh corresponding to that subject's neutral face. Then all the subjects are aligned together using the same method.

Points that are known to be highly deformable (for example, on the mouth) are weighed less in the transformation. In our experiments, we set the weight to 0.9 for the eyes and nose and 0.1 for the lips and eyebrows. It is our observation that the alignment, and shape learning, is not very sensitive to the weights.

2.4. Model Computation

The aligned tracked meshes are used to compute the model mean shape $\overline{\mathbf{X}}^i$, shape matrix $\mathbf{S}^i$ and action matrix $\mathbf{A}^i$. Principal Component Analysis is used.

The mean shape is computed as the mean of the neutral expression meshes $\mathbf{X}_{N,s}$ of all the subjects $1, \ldots, N_S$. Equal weighting is given to each subject.

$$\overline{\mathbf{X}} = \frac{1}{N_S}\sum_{s=1}^{N_S}\mathbf{X}_{N,s} \tag{4}$$

Next, the mean shape is subtracted from the neutral expression shapes of all the subjects. We apply Principal Component Analysis on the matrix each column of which contains the resulting shape $(\mathbf{X}_{N,s} - \overline{\mathbf{X}})$ for each subject and keep the first $p = 2$ principal vectors as the shape basis vectors in the matrix $\mathbf{S}$.

Then, the mean shape is subtracted from all the meshes of the subjects, and components orthogonal to the shape basis vectors are computed. The result is assembled in a matrix and principal component analysis is used to compute the action basis vectors from this matrix. The most significant $q = 5$ vectors are kept in the action matrix $\mathbf{A}$.

We find the components of the original aligned meshes orthogonal to the action basis vectors just calculated. Having removed the action component from the meshes, the resulting meshes are used as neutral expression meshes to re-estimate the Mean shape and the shape matrix as described earlier. The action matrix is then recalculated and the process is iterated until the solution converges. The algorithm converges in about 10–20 iterations.

2.5. Monocular Tracking

Optical flow tracking computes the translational displacement of points in the image given two successive frames [14, 15]. Each image point is processed independently. In model-based tracking, the points are constrained by the parameterized model (see equation 3). So, the parameters are estimated simultaneously from image measurements, similar to [12].

The shape parameter vector σ is initialized only at the start of the sequence and is not allowed to vary later.

Assume that the face model has been tracked from the first frame to the $(n-1)$th frame I_{n-1}. The objective is to estimate the optimal pose ($\omega(n)$, $\mathbf{t}(n)$) and action parameters ($\alpha(n)$) that best fit the subsequent frame I_n. The cost function to be minimized for the purpose is given by

$$C_n = \sum_{i,ROI}\left\{\begin{array}{l}(1-\epsilon)(I_n(\mathbf{x}^i(n)) - I_{n-1}(\mathbf{x}^i(n-1)))^2 \\ +\epsilon(I_n(\mathbf{x}^i(n)) - I_1(\mathbf{x}^i(1)))^2\end{array}\right\} \tag{5}$$

Observe that the first term in equation 5 is the standard matching cost used in the Shi-Tomasi-Kanade tracker [14, 15]. The second term measures the image mismatch between the current image I_n and the first image I_1. This additional term weakly forces the facial features to appear the same over the complete sequence.

Hence, it avoids tracking drift and increases robustness. It is referred to as the drift monitoring term.

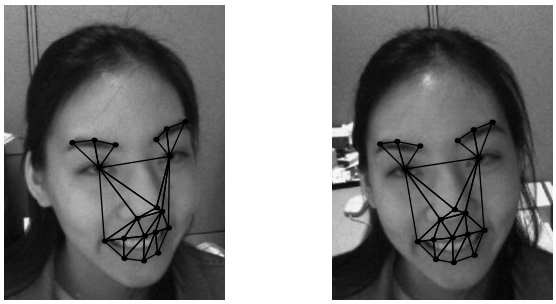
3. EXPERIMENTAL RESULTS

In this section, we present and discuss experimental results for stereo tracking, shape learning and monocular model-based tracking.

For data acquisition, we used the Digiclops^{TM} camera system [17] at an average frame rate of 10fps, with color images of size 640×480. We acquired stereo image sequences for 11 subjects (8 male, 3 female). Four sequences per subject were recorded: two for the training set, without any rigid motion of the head; and two for testing where the subject could move his head freely. Each of the sequences was 380 frames long.

3.1. Stereo Tracking

Stereo tracking was performed on the sequences in which the subject exhibits various facial expressions without changing his or her pose. Initialization of these sequences was done manually. As discussed in Section 2.2, the tracking algorithm as well as the tuning parameters were the same as in [12]. The quality of the tracking was judged visually based on the reprojection of the mesh points on the left and right camera images. Figure 2 shows the tracking results for one image pair. The maximum reprojection error of the mesh onto the left and right images was estimated to be less than 2 pixels.



(a) Left image (b) Right image

Fig. 2. Stereo sequence used for shape learning.

3.2. Shape Learning

The stereo tracking data was used to learn the model (given by equation 3) as described in Section 2.4. All the subjects' data was used in the computation of the model. To find out if the computed model covers the space of the shapes in our training set, we calculated the 2-dimensional shape parameter σ for all the frames of the subjects. The plot in σ-space is shown in Figure 3. It can be seen that each subject's shape parameter points are somewhat clustered together. This shows that the shape vector represents the differences in shape between people in some form. It should also be noted that the computed shape parameters can be useful for recognizing faces.

In another experiment, we calculated the error between the neutral face shape of every subject and its approximation by the model where that subject was not included in the training. The rms error of the model points with respect to the actual mesh points was 3.53mm. Figure 4 shows the projection of the models on the left and right images of two subjects (circles) and the actual mesh points (shown as plus signs).

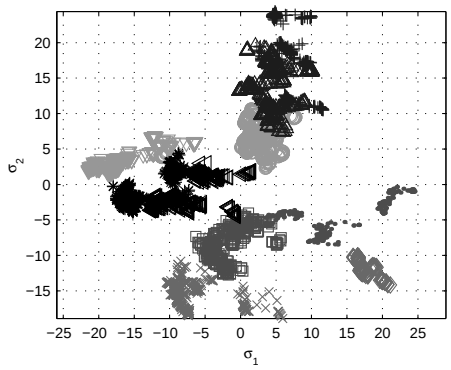


Fig. 3. Projections of all the meshes in σ space.

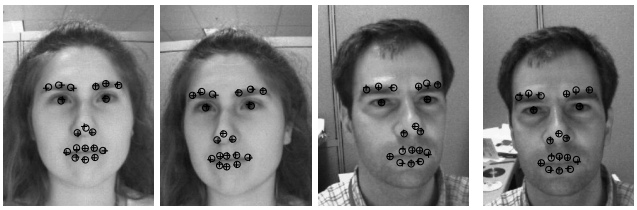


Fig. 4. Model estimates for a few subjects. The $+$ is for model approximation and the circle for actual mesh points.

To check robustness, we excluded one subject at a time from the training set and computed the model in each case. There is no obvious method to measure similarity of two linear subspaces. For our task, we used the inner product as a measure of the similarity of the basis vectors. Since the basis vectors are unit magnitude, therefore inner product being equal to ± 1 means the vectors are equivalent. We calculated the similarity measure between the "all subjects in training set" model and each of the models where one subject was excluded. The average similarity measure for the shape matrix $\mathbf{S}$ and the action matrix $\mathbf{A}$ were 0.988 and 0.948 respectively while the mean shape $\overline{\mathbf{X}}$ varied by an average of 1.7%.

3.3. Monocular Tracking

The model was initialized semi-automatically on the first frame of the monocular sequence. Then the monocular tracking algorithm, discussed in Section 2.5, was used to compute the pose (ω and $\mathbf{t}$) and the action parameter vector α for every frame. The shape parameter vector σ was computed at initialization and was not updated during tracking. The value of the drift monitoring coefficient was set to $\epsilon = 0.2$ for all our experiments to emphasize standard tracking cost over drift monitoring cost.

Figures 5 and 6 show images from two test sequences each consisting of 380 frames. The subjects being tracked were not included in the training data set for model computation. Throughout the sequence, the subject rotates and moves his head covering a

working volume of $10cm \times 10cm \times 10cm$ while doing a variety of facial expressions. It can be seen that the tracking of the face is maintained over the length of the sequence.

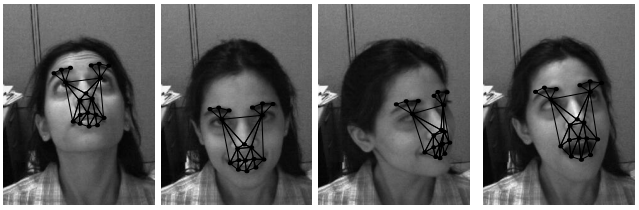


Fig. 5. Monocular tracking results on subject 1.

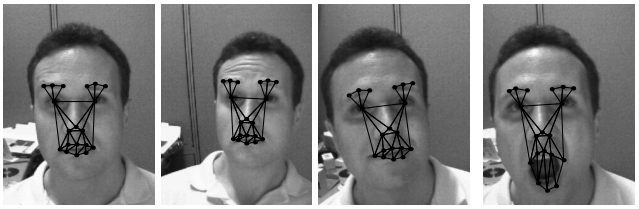


Fig. 6. Monocular tracking results on subject 2.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented a system for building a simple 3D model of the face from real data and use of that model for 3D tracking of the pose and action deformations of the human face in monocular image sequences without the use of special markers. There are three main contributions of this paper. First, we demonstrated that a data-driven approach for model construction offers an elegant and practical alternative to the task of manual construction of models using 3D scanners or CAD modelers. Second, we showed that both shape and action deformations can be learned simultaneously from real data. Third, we demonstrated that the learned model is robust and generalizes well to subjects not included in the training set. We also showed that this model gives good results for face tracking in monocular image sequences.

There are many possible directions for future work. The use of shape parameter vector to recognize faces can be investigated. We also intend to work on a real-time implementation of the tracking algorithm. For that purpose, we need to develop an automatic initialization procedure. Further investigation into the dimensionality of shape and action spaces is also planned. It will also be interesting to extend this model of face geometry to a model using both geometry and appearance.

5. REFERENCES

[1] D. DeCarlo and D. Metaxas, "The integration of optical flow and deformable models with applications to human face shape and motion estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, June 1996, pp. 231–238.

[2] R. Okada, Y. Shirai, and J. Miura, "Tracking a person with 3-D motion by integrating optical flow and depth," in *Proc. of the 4th IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, March 2000, pp. 336–341.

[3] S.-C. Pei, C.-W. Ko, and M.-S. Su, "Global motion estimation in model-based image coding by tracking three-dimensional contour feature points," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 8, pp. 181–190, April 1998.

[4] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for computer vision," Draft report, Imaging Science and Biomedical Engineering, University of Manchester, UK, September 2001, http://www.isbe.man.ac.uk.

[5] J. Ahlberg, "Using the active appearance algorithm for face and facial feature tracking," in *Proc. of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, July 2001, pp. 68–72.

[6] D. DeCarlo and D. Metaxas, "Deformable model-based face shape and motion estimation," in *Proc. of the 2nd Intl. Conf. on Automatic Face and Gesture Recognition*, October 1996, pp. 146–150.

[7] P. Eisert and B. Girod, "Analyzing facial expressions for virtual conferencing," *IEEE Computer Graphics and Applications*, vol. 18, pp. 70–78, September 1998.

[8] P. Eisert and B. Girod, "Model-based estimation of facial expression parameters from image sequences," in *Proc. of the IEEE Intl. Conference on Image Processing*, October 1997, vol. 2, pp. 418–421.

[9] J. Ahlberg, "CANDIDE-3 – an updated parametrized face," Tech. Rep. LiTH-ISY-R-2326, Dept. of EE, Linköping University, Sweden, January 2001.

[10] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, June 2000, pp. 690–696.

[11] Volker Blanz and Thomas Vetter, "Morphable model for the synthesis of 3-D faces," in *Proc. of SIGGRAPH'99*, Aug 1999, pp. 187–194.

[12] S. B. Gokturk, J.-Y. Bouguet, and R. Grzeszczuk, "A data-driven model for monocular face tracking," in *Proc. of the IEEE Intl. Conf. on Computer Vision*, July 2001, vol. 2, pp. 701–708.

[13] Olivier Faugeras, *Three dimensional computer vision*, MIT Press, 1993.

[14] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of the 7th Intl. Joint Conf. on Artificial Intelligence*, 1981, pp. 674–679.

[15] J. Shi and C. Tomasi, "Good features to track," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, June 1994, pp. 593–600.

[16] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistal Society B*, pp. 285–339, 1991.

[17] Digiclops Stereo System. Point Grey Research., Visit `http://www.ptgrey.com/products/digiclops`.