3D Face Pose Tracking From an Uncalibrated Monocular Camera

Zhiwei Zhu
Department of ECSE, RPI, Troy, NY
zhuz@rpi.edu

Qiang Ji
Department of ECSE, RPI, Troy, NY
qji@ecse.rpi.edu

1 Abstract

We propose a new near-real time technique for 3D face pose tracking from a monocular image sequence obtained from an uncalibrated camera. The basic idea behind our approach is that instead of treating 2D face detection and 3D face pose estimation separately, we perform simultaneous 2D face detection and 3D face pose tracking. Specifically, 3D face pose at a time instant is constrained by the face dynamics using Kalman Filtering and by the face appearance in the image. The use of Kalman Filtering limits possible 3D face poses to a small range while the best matching between the actual face image and the projected face image allows to pinpoint the exact 3D face pose. Face matching is formulated as an optimization problem so that the exact face location and 3D face pose can be estimated efficiently. Another major feature of our approach lies in the use of active IR illumination, which allows to robustly detect eyes. The detected eyes can in turn constrain the face in the image and regularize the 3D face pose, therefore the tracking drift issue can be avoided and the processing can speedup. Finally, the face model is dynamically updated to account for variations in face appearances caused by face pose, face expression, illumination and the combination of them.

2 Introduction

Face pose tracking is very important in vision based applications such as Human-Computer Interaction (HCI), Face Recognition, and Virtual Reality. Many techniques have been proposed for face pose estimation. Basically, face pose estimation techniques can be classified into two main categories: appearance-based approaches [1, 2, 3] and model-based approaches [4, 5, 6]. Appearance-based approaches attempt to use holistic facial appearance, where face is treated as a two-dimensional pattern of intensity variations. They assume that there exists a mapping relationship between 3D face pose and certain properties of the facial image, which is constructed from a large number of training images. Model-based (or features-based) approaches usually assume a 3D model of the face and recover the face pose based on the assumed model. First, a set of $2D - 3D$ feature correspondences are established. Then the face pose is estimated by using the conventional pose estimation techniques. So, pose estimation is done after face feature tracking.

We can conclude that most existing methods for face pose estimation follow the strategy of face detection in the image and face pose estimation from the detected face images. The main problem with all these approaches is that face detection and face pose estimation are carried out independently. There is no input from each other. But in reality, these two steps are very interrelated, and because the face location in the image is caused by face pose in 3D, they must be consistent with each other. Therefore, we propose to take full advantage of the interdependent relationship between face image and 3D face pose and perform face detection and face pose estimation simultaneously.

In this paper, we describe a new technique to perform the 2D face tracking and 3D pose estimation synchronously. In our method, 3D face pose is tracked using Kalman Filtering. The initial estimated 3D pose is then used to guide face tracking in the image, which is subsequently used to refine the 3D face pose estimation. Face detection and face pose estimation work together and benefit from each other. Weak perspective projection model is assumed so that face can be approximated as a planar object with facial features, such as eyes, nose and mouth, located symmetrically on the plane. Figure 1 summarizes our approach. First, we automatically detect a frontal-parallel face view image based on the detected eyes and some simple anthropometric statistics. The detected face region is used as the initial 3D planar face model. The 3D face pose is then tracked starting from the frontal-parallel face pose. During tracking, the 3D face model is updated dynamically to account for the face appearance changes introduced by the illumination, facial expression, face pose, or the combinations of them, and the face detection and face pose estimation are synchronized and kept consistent with each other. Also, the use of the 2D eye location is used to constrain the face location in the image, which can avoid the tracking drift issue and speedup the processing.

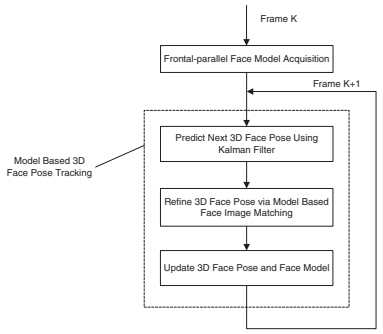


Figure 1. The flowchart of face pose tracking

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

3 Weak Perspective Model for Face Pose Estimation

We employ an object coordinate system affixed to user's face, with face normal (face pose) being the Z axis of the object frame. Without loss of generality, face is assumed to be planar. Let $X = (x, y, 0)^t$ be the coordinate of a 3D point on the face relative to the object coordinate frame, and $p = (c, r)^t$ the coordinate of the corresponding projection image point in the row-column frame.

For each pixel (c_1, r_1) in a given image frame I_1 and the corresponding image point (c_2, r_2) in another image I_2, using basic projection equation of weak perspective camera model for planar 3D object points [7] yields

$$\begin{pmatrix} c_2 - c_{c2} \\ r_2 - r_{c2} \end{pmatrix} = M_2 * {M_1}^{-1} \begin{pmatrix} c_1 - c_{c1} \\ r_1 - r_{c1} \end{pmatrix} \quad (1)$$

where M_1 and M_2 are the projection matrices for image I_1 and I_2 respectively, and (c_{c1}, r_{c1}) and (c_{c2}, r_{c2}) are the projection points of the same reference point (x_c, y_c, z_c) in the image I_1 and image I_2. Equation 1 is the fundamental weak perspective homographic projection equation that relates image projections of the same 3D points in two images with different face poses. The homographic matrix $P = M_2 * {M_1}^{-1}$ characterizes the relative orientation between the two face poses.

The face pose can be characterized by a rotation matrix R resulted from successive Euler rotations of the camera frame around its X axis by ω, its once rotated Y axis by ϕ, and its twice rotated Z axis by κ [7]. Elements of the projection matrix is related with these three angles and a scale factor λ, representing the distance from face to camera. The 3D pose of a face can therefore be characterized by the three Euler angles and the scaler λ. While the three angles determine face orientation, λ determines the distance from face to camera. Face pose estimation can be expressed as determination of these four parameters.

4 Simultaneous 3D Face Pose Determination and 2D Face Tracking

We propose a novel technique to track 3D face pose and 2D face in the image synchronously as follows.

4.1 Automatic 3D Face Model and Pose Initialization

In our algorithm, we should have a fronto-parallel face to represent the initial face model. This initialization is automatically accomplished by using the eye tracking technique we have developed. Specifically, the subject starts in fronto-parallel face pose position with the face facing directly to the camera as shown in figure 2. The eye tracking technique is then activated to detect eyes. After detecting the eyes, the first step is to compute the distance d_{eyes} between two eyes. Then, the distance between the detected eyes, eyes locations and the anthropometric proportions are used to estimate the scope and the location of the face in the image automatically. Experiments show that our face detection method works well for all the faces we tested. Example of the detected frontal face region is shown in Figure 2. Once the face region is decided, we will treat it as our initial face model, whose pose parameters are used as initial face pose parameters.

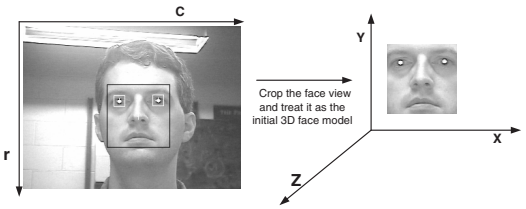


Figure 2. The initial face model

The face pose parameters $(\omega, \phi, \kappa, \lambda)$ for the initial face model are $(0°, 0°, 0°, 1)$, where λ, without loss of generality, is normalized to 1. The corresponding projection matrix M is:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (2)$$

We use the center of the 3D face model as the reference point, which is represented as (x_c, y_c, z_c). Furthermore, we assume its corresponding projection image point is (c_c, r_c) in the row-column image plane.

4.2 Face Pose Tracking Algorithm

4.2.1 Kalman Filter with Eye Constraints

Kalman filter is a well-known prediction method for tracking. Given the 2D face model obtained from the initialization, the current face pose parameters $X_t = (\omega_t, \phi_t, \kappa_t, \lambda_t)$ and the current face image location (x_t, y_t), the Kalman Filtering for face pose tracking can be modelled nicely according to the theory of Kalman Filtering [8]. The use of Kalman Filtering limits possible 3D face poses to a small range by putting a smooth constraint on each of the six face pose parameters.

But the prediction based on Kalman Filtering will be off significantly if head undergoes a sudden rapid movement. In dealing with the problem, we propose to approximate the face movement with eyes movement since eyes can be reliably detected in each frame. Let the predicted face pose vector at $t + 1$ based on eyes motion be X^p_{t+1}. Then the final predicted face pose should be based on combining the one from Kalman X^-_{t+1} with the one from eyes, i.e.,

$$X^*_{t+1} = X^-_{t+1} + \Sigma^{-1}_{t+1}(X^-_{t+1} - X^p_{t+1}) \quad (3)$$

where Σ^{-1}_{t+1} is the uncertainty at time $t + 1$.

The simultaneous use of Kalman Filtering and eyes motion allows to perform accurate face pose prediction even under significant and rapid head movements. We can then derive a new covariance matrix Σ^*_{t+1} for X^*_{t+1} accordingly to characterize its uncertainty.

4.2.2 Face Detection and Pose Estimation via Matching Optimization

The combined prediction from Kalman Filtering provides the predicted face position, 3D face pose, and the associated

uncertainty Σ^*_{t+1} for the next frame $t+1$. Σ^*_{t+1} can be used to limit the search area for the face location and face pose at time $t+1$. Face detection and face pose estimation are to search for a face position and 3D face pose within the scope determined by Σ^*_{t+1} such that the detected face view image can best match the projected face view image under the given face pose.

Mathematically, this is formulated as follows. Find the state vector z_{t+1} within the scope determined by Σ^*_{t+1} from X^*_{t+1} such that the detected face image best matches the projected face image. We formulate the matching criterion as Sum of Squared Difference errors over all the image pixels within the region of interest:

$$E_{matching} = \sum_{i=1}^{N} (I_p(i) - I_c(i))^2 \quad (4)$$

where, $I_p(i)$ is the pixel value of ith pixel in the reconstructed face view image I_p, $I_c(i)$ the pixel value of the face image in the current image frame, and N the total pixel number of the reconstructed face view image. $I_p(i)$ is projected from the reference image $I_r(c,r)$ via a mapping function f

$$I_p = f(I_r, \alpha) \quad (5)$$

where $\alpha = (\omega, \phi, \kappa, \lambda, x, y)$, which consists of the face pose parameters. We need to find the locally optimal face pose parameter set ${\alpha^*}_n$, which results in the projected face image that best matches the face in the current image frame

$${\alpha^*}_n = \arg\min_{\alpha} E_{matching}$$

$E_{matching}$ is minimized to solve for the 6 pose parameters, where x^*_{t+1} serves as the initial value of α.

4.2.3 Regularization of Minimization Criteria by Eyes Positions

Since different sets of pose parameters may produce the same image, yielding the same $E_{matching}$, the minimization procedure may converge to a wrong place if it is left unconstrained. This problem is further exacerbated by the susceptibility of SSD to drifting. To overcome this, a penalty term is imposed to each SSD error corresponding to each set of pose parameters. Since we can accurately and independently detect the eyes position, the detected eyes positions can be used to constrain the 3D face pose and the 2D face image.

For each pair of the projected eyes in the projected face image and the detected eyes in the detected face image, the distance E_{eyes} between the detected eyes and the projected eyes can be expressed as

$$E_{eyes} = E_{Left} + E_{Right} \quad (6)$$

where E_{Left} is the Euclidean distance between the detected left eye and the projected left eye, and E_{Right} is the Euclidean distance between the detected right eye and the projected right eye. The correct face pose and face position should simultaneously minimize the results from equations 4 and 6. Therefore, the criteria for the image matching minimization can be expressed as the sum of both terms

$$E = \beta E_{matching} + (1-\beta) E_{eyes} \quad (7)$$

where β is a scale factor determining the relative importance of two terms. It is determined empirically.

4.2.4 Dynamic Face Model Updating

In practice, during tracking, the human face usually undergoes different kinds of variations, most of which come from the pose, expression, illumination and the combination of them. In order for face templates to maintain adequate tracking performance while tracking face with time-varying appearance, it is necessary to dynamically adapt the face model to keep them consistent with the changing appearance of the face.

Specifically, our strategy for updating the face template is to include in each new face template with a portion of the initial template, which is assumed to have been chosen correctly, and thus contain the desired target. For example, for the k^{th} frame, once the best match is found, with the matching confidence measurement above a certain threshold, the new face template $I_r(k+1)$ for frame $k+1$ is updated to be the combination of the initial face template $I_r(0)$ and the face model $I_g(k)$ generated from the image region $I_c(k)$ in the new frame that best matches the current template as follows:

$$I_r(k+1) = \rho I_r(0) + (1-\rho) I_g(k)$$

where $I_g(k) = g(I_c(k), \alpha)$, and α is the detected face pose parameters related with frame k and function g is the inverse projection function of equation 5.

Furthermore, the constant $\rho \in [0,1]$ determines the contribution of the initial template to the new template. $\rho = 0$ is the case of fully updated templates and $\rho = 1$ gives standard templates.

By the dynamic face model updating and physical eye locations constraints, our tracker greatly improve the tracking accuracy.

5 Experimental Results

A series of experiments involving real image sequences are conducted to characterize the performance of our face pose estimation technique. First, the accuracy of the estimated face pose is analyzed. Then, we study the sensitivity of our algorithm to perturbations with initial face pose and placement.

5.1 Accuracy of the Face Pose Tracker

In order to measure the accuracy of the estimated face pose, several image sequences were collected. The ground truth for these sequences was simultaneously collected via "Flock of Birds" magnetic tracker.

In figure 3, two estimated face pose angles, Pitch and Yaw, are shown together with the ground truth. The estimation error for Yaw is 2.9260 degrees and the estimation error for Pitch is 3.6174 degrees.

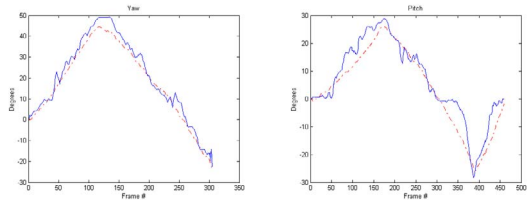
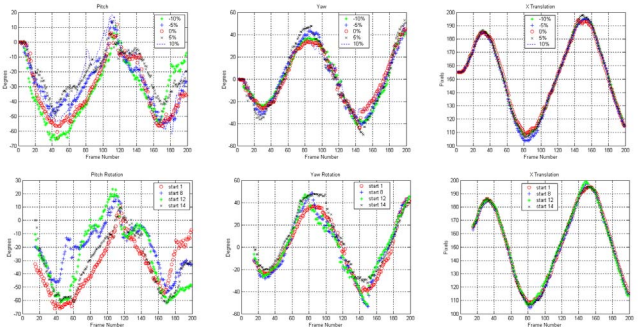


Figure 3. Comparison between the estimated face pose and the ground truth. In each graph, the dashed line depicts the ground truth and the solid line depicts the estimated face pose.

5.2 Sensitivity of the Face Pose Tracker

Two experiments are conducted to test sensitivity of the tracker to the initial face size and the initial face pose for the face model separately. One sequence that contains a person rotating his face naturally in front of the camera is chosen for this experiment. First, we perturb the size of face model, which is represented by the face width, by ± 5 and ± 10 percent of the initial face width. Figure 4 shows some tracking results corresponding to the different variations of face sizes. We can see that the tracking trajectories basically follow the same trend though there are some small deviations at certain frames. Second, we simulate the perturbations to the frontal-parallel pose angles of the face model by choosing the initial face model from the image frame, which contains the face not under the frontal-parallel pose. Then the tracking is started from that image frame. We choose frames 1, 8, 12 and 14 as the starting frame for tracking respectively. They roughly correspond to the following face poses: $(0°, 0°, 0°)$, $(-0.4°, -0.5°, -1°)$, $(-8.8°, -10.9°, -2°)$ and $(-11.7°, -12.8°, -2.7°)$. Figure 4 shows some tracking results.



Pitch rotation Yaw rotation X Translation

Figure 4. (a) First row: sensitivity of the tracker to errors in estimating size of the initial face model. (b) Second row: sensitivity of the tracker to errors in the initial face pose of face model.

We can conclude that the tracker is not very sensitive to slight variations of initial face poses.

5.3 Convergence and Speed

Currently, our system can reach about 10 fps in a computer with an Intel Pentium 4 processor and $512M$ memory. Further speed gain can be obtained with a faster computer and additional optimization of the codes. Some tracking results of our system are shown in Figure 5. Video demos of our system may be found at http://www.ecse.rpi.edu/~cvrl/Demo/demo.html.

Figure 5. Face pose tracking results taken from one video experimental sequence, which are randomly selected. The white rectangle indicates the tracked face region and the white line is the norm of the face plane which is drawn according to those three angles.

6 Conclusions

In this paper,we present a technique for simultaneous 3D face pose tracking and face detection under different face orientations, facial expressions and illumination changes. Experimental results show how our technique greatly improves the standard pure planar face tracker, but still enjoys its simplicity. Based on the proposed techniques, we have built a real time working system that will start to track the face and estimate the 3D face pose as soon as the user is sitting in front of the camera.

References

[1] S. McKenna and S. Gong, "Real-time face pose estimation," *Int. J. Real Time Imaging, Special Issue on Visual Monitoring and Inspection*, vol. 4, pp. 333–347, 1998.

[2] M. Motwani and Q. Ji, "3d face pose discrimination using wavelets," in *ICIP*, October 7-10,2001.

[3] S. Nayar, H. Murase, and S. Nene, "Parametric appearance repesentation," in *In Early Viual Learning. Oxford University Press*, 1996.

[4] P. Yao, G. Evans, and A. Calway, "Using affine correspondence to estimate 3-d facial pose," in *ICIP*, 2001, pp. 919–922.

[5] A. Gee and R. Cipolla, "Determining the gaze of faces in images," *Image and Vision Computing*, vol. 12, pp. 639–948, 1994.

[6] T. Horprasert, Y. Yacoob, and L.S. Davis, "Computing 3-d head orientation from a monocular image sequence," in *International Conference on AFGR*, 1996.

[7] E. Trucco and A. Verri, "Introductory techniques for 3-d computer vision," *Prentice-Hall, New Jersey*, 1998.

[8] Kalman, Rudolph, and Emil, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.