3D Facial Pose Tracking in Uncalibrated Videos*

Gaurav Aggarwal, Ashok Veeraraghavan, and Rama Chellappa

University of Maryland, College Park MD 20742, USA,
{gaurav, vashok, rama}@umiacs.umd.edu
WWW home page: http://www.cfar.umd.edu/~gaurav

Abstract. This paper presents a method to recover the 3D configuration of a face in each frame of a video. The 3D configuration consists of the 3 translational parameters and the 3 orientation parameters which correspond to the yaw, pitch and roll of the face, which is important for applications like face modeling, recognition, expression analysis, etc. The approach combines the structural advantages of geometric modeling with the statistical advantages of a particle-filter based inference. The face is modeled as the curved surface of a cylinder which is free to translate and rotate arbitrarily. The geometric modeling takes care of pose and self-occlusion while the statistical modeling handles moderate occlusion and illumination variations. Experimental results on multiple datasets are provided to show the efficacy of the approach. The insensitivity of our approach to calibration parameters (focal length) is also shown.

1 Introduction

Face tracking is a crucial task for several applications like face recognition, human computer interaction, etc. Most of these applications require actual 3D parameters of the location of the head. In this paper, we propose an approach based on a cylindrical model for the head for reliable tracking of position and orientation of the head under illumination changes, occlusion and extreme poses.

There has been significant work on facial tracking using 2D appearance based models[1][2][3][4]. Quite clearly, such 2D approaches do not explicitly provide the 3D configuration of the head. Recently, several methods have been developed for 3D face tracking. [5] uses a closed loop approach that utilizes a structure from motion algorithm to generate a 3D model of the face. In [6], techniques in continuous optimization are applied to a linear combination of 3D face models. [7] proposes a hybrid sampling solution using both RANSAC and particle filters to track the pose of a face. [8] shows examples of head tracking by posing it as a nonlinear state estimation problem. A cylindrical face model for face tracking has been used in [9]. In their formulation, they assume that the inter-frame warping function is locally linear and that the inter-frame pose change occurs only in one of the six degrees of freedom of the rigid cylindrical model. In our approach, we do not make any such assumptions. This improves both tracking accuracy and robustness. Moreover, unlike [9] we do not use information about

* Partially supported by the NSF-ITR Grant 03-25119

the camera calibration parameters. Instead we analytically show the insensitivity of our approach to errors in focal length.

2 The Geometric Model

The choice of the model to represent the facial structure is very crucial for the problem of face tracking. There are several algorithms that do not assume an explicit structural model but track salient points, features or 2D patches [10] [11][4][2]. On the other extreme, there are algorithms like [5] that use a set of 3D laser-scanned heads. Though a planar model will probably be the simplest one to use, it does not have the capability to estimate the 3D configuration of the face. On the other hand, using a complicated face model makes the initialization and registration process difficult.

Similar to [9], we use a cylindrical model, though with an elliptical cross-section, to represent a face. Assuming that our cylindrical model reasonably approximates the 3D structure of a face, the problems related to pose and self-occlusion get automatically taken care of. Due to the absence of the calibration parameters, people usually assume orthographic projection. The use of orthographic projection is restrictive and introduces confusion between scale and pitch. These reasons motivate us to use perspective projection model. Since we do not know the camera focal length for un-calibrated videos, we show that our approach for pose recovery is robust to the errors in focal length assignment.

Let us assume that the true focal length of the camera imaging a cylinder centered at (X_0, Y_0, Z_0) with height H and radius R be f_0. Let us assume that we erroneously set the focal length to kf_0 (without loss of generality $k \geq 1$). The true projections of feature points on the cylinder are given by

$$x_f = \frac{f_0 X_f}{Z_0 + z_f} \qquad y_f = \frac{f_0 Y_f}{Z_0 + z_f} \qquad \text{where,} \qquad Z_f = Z_0 + z_f \tag{1}$$

The projection of feature points of another cylinder with same dimensions but placed at (X_0, Y_0, kZ_0) and imaged by a camera of focal length kf_0 are

$$\hat{x}_f = \frac{kf_0 X_f}{kZ_0 + z_f} = x_f \left[1 + \frac{(k-1)z_f}{kZ_0 + z_f}\right] = x_f[1 + \delta_f] \tag{2}$$

$$\hat{y}_f = \frac{kf_0 Y_f}{kZ_0 + z_f} = y_f \left[1 + \frac{(k-1)z_f}{kZ_0 + z_f}\right] = y_f[1 + \delta_f] \tag{3}$$

If $\delta_f \ll 1$, the feature positions for the cylinder at (X_0, Y_0, Z_0) imaged by camera f_0 is equivalent to a cylinder at (X_0, Y_0, kZ_0) imaged by a camera with focal length kf_0. Therefore, when δ_f is small, our estimates of yaw, pitch and roll are reasonably accurate.

If the depth variations in the object (cylinder in our case) are smaller than the distance of the object from the camera center (i.e., $z_f \ll Z_0$) and the field of view is reasonably small, then

$$\delta_f = \frac{(k-1)z_f}{kZ_0 + z_f} \quad < \quad \frac{kz_f}{kZ_0 + z_f} \quad < \quad \frac{\frac{z_f}{Z_0}}{1 + \frac{z_f}{kZ_0}} \quad \ll \quad 1 \tag{4}$$

The choice of features is extremely important for the task of 3D pose estimation of a moving face. The features should be easy to detect, robust to occlusions, changes in pose, expression and illumination. In this paper, we propose a hybrid approach which makes use of the advantages of a purely geometric approach and the power of statistical inference. We use an extremely simple and easily computable feature to stress-test the approach. We superimpose a rectangular grid all around the curved surface of our elliptical cylinder. The mean intensity for each of the visible grids forms the feature vector. Robust statistics makes the feature robust to moderate illumination changes, expressions and occlusions.

3 Tracking Framework

Once the structural model and feature vector have been fixed, the goal is to estimate the configuration (or pose) of the moving face in each frame of a given video. This can be viewed as a dynamic state estimation problem. Here the state consists of the six 3D configuration parameters. Particle filtering [12][13] is an inference technique for estimating the unknown dynamic state θ of a system from a collection of noisy observations $y_{1:t}$. The two components of this approach are the state transition model which models the state evolution, and the observation model which specifies the state-observation dependence:

$$\text{State transition model: } \theta_t = f(\theta_{t-1}, u_t), \tag{5}$$

$$\text{Observation model: } y_t = g(\theta_t, v_t), \tag{6}$$

where u_t is the system noise while v_t is the observation noise. In general, the functions f and g can also be time-dependent. The particle filter approximates the desired posterior pdf $p(\theta_t|y_{1:t})$ by a set of weighted particles $\{\theta_t^{(j)}, w_t^{(j)}\}_{j=1}^N$, where N denotes the number of particles. The state estimate $\hat{\theta}_t$ is recovered from the pdf as the maximum likelihood (ML) estimate. To keep the tracker as generic as possible, we use a random-walk model as the motion model:

$$\theta_t = \theta_{t-1} + u_t, \tag{7}$$

where u_t is normally distributed about zero. Based on the domain knowledge, one can come up with a motion model that will be capable of estimating the pdf better with lesser number of particles.

The observation model involves the feature vector described in the previous section. In our framework, we can rewrite the observation equation as:

$$z_t = \Gamma\{y_t; \theta_t\} = F_t + v_t, \tag{8}$$

where y_t is the current frame (the gray scale image), Γ is the mapping that computes the feature vector given an image y_t and a configuration θ_t, z_t is the computed feature vector and F_t is the feature model. The feature model is used to compute the likelihood of the particles (which correspond to different proposed configurations of the face). For each particle the likelihood is computed using

the average sum of squared differences (SSD) between the feature model and the *mean vector* z_t corresponding to the particle.

On one extreme, the feature model can be a fixed template, while on the other hand one can use a dynamic template e.g, $F_t = \hat{z}_{t-1}$. Similar to [14], we refer to the fixed template $F_t = F_0$ as the lost model while the dynamic component $F_t = \hat{z}_{t-1}$ as the wander model. It is worthwhile to note that the wander component is capable of handling appearance changes due to illumination, expression, etc. as the face translates/rotates in the real world, while the lost component is resistant against drifts. In the current implementation, the likelihood of a particle is computed as the maximum of the likelihoods using the lost and the wander model. This gives us the capability both to handle appearance changes and to correct the estimation if the wander model drifts. The tracker performs well with as few as 200 particles. The performance does not show any appreciable improvement as we increased this number.

We use robust statistics in our likelihood model in order to make the feature robust to changes in illumination, expression and occlusions. We trust only the top half of the means and treat the rest as outliers as follows:

$$p(y_t|\theta_t^{(j)}) = e^{-\lambda dist} \quad \text{where,} \quad dist = \frac{\sum_{m,n} \eta(m,n) d(m,n)}{\sum_{m,n} \eta(m,n)} \tag{9}$$

where $\eta(m,n)$ is the visibility indicator variable, while $d(m,n)$ is computed as:

$$d(m,n) = \begin{cases} (F_t(m,n) - z_t^{(j)}(m,n))^2 & \text{if } d(m,n) < c \\ c & \text{otherwise} \end{cases}$$

$$\text{where,} \quad c = median(\{d(m,n)\}) \tag{10}$$

4 Experimental Results and Conclusion

Tracking under extreme poses: We conducted tracking experiments on 3 datasets (Honda/UCSD dataset [15], BU dataset [9] and Li dataset [16]). These datasets have numerous sequences in which there are significant illumination changes, expression variation and people are free to move their heads arbitrarily. Figure 1 shows few of the frames from three videos with grid points on the estimated cylinder overlaid on the image frame. The first and second columns show the ability of the tracker to maintain tracks in spite of considerable occlusion. The tracker does well even when confronted with extreme poses as shown in the third column. Moderate expressions do not affect the feature since it is the mean intensity within a small surface patch on the face. During certain severe expression changes robust statistics helps maintain the track. The tracker is able to maintain the track all along the sequences.

Ground Truth Comparison: The BU dataset [9] provides the ground truth for the pose of the face in each frame. We conducted tracking experiments on the BU dataset and compared the yaw, pitch and roll estimated by the tracker

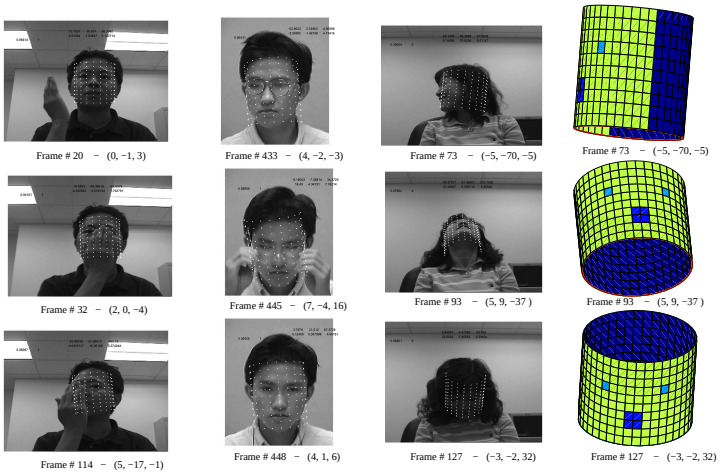

Fig. 1. Tracking results under severe occlusion, extreme poses and different illumination conditions. The cylindrical grid is overlaid on the image plane to display the results. The 3-tuple shows the estimated orientation (roll, yaw, pitch). The last column shows the cylindrical model in the pose estimated for the sequence in the third column.

to the ground truth. Figure 2 shows the comparison for three sequences. We see that the tracker accurately estimates the pose of the face in most frames.

Recognition with non-overlapping poses: Most recognition methods require the gallery and probe images to be in similar pose. Since the tracking method maintains explicit pose for each frame, we do not need this. We show this by performing recognition on non-overlapping poses. The closest poses in the gallery and the probe differ by at least 30 degrees. We used 10 subjects from the Honda/UCSD dataset [15] for this experiment. For each frame we build a texture mapped cylinder using the tracked pose. We used the minimum sum of squared distance between a gallery model and a probe model as the distance between two videos. This is a very challenging experiment since the poses exhibited by the gallery videos and those exhibited by the probe videos are very different. In spite of this, we obtained 100% recognition rate in this experiment.

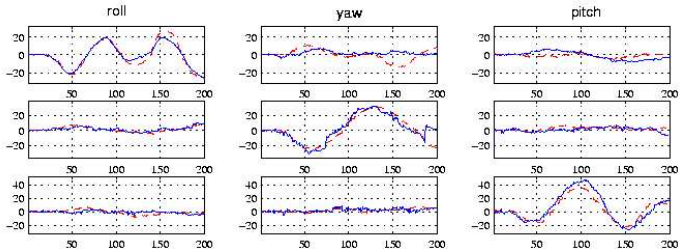

Fig. 2. Each row shows the 3 orientation parameters. The red/dashed curve depicts the ground truth while the blue/solid curve depicts the estimated values.

Conclusions: We have proposed a method for tracking the facial pose in a video. The tracker is robust to moderate occlusions and illumination changes

and maintains track even during extreme poses. We have also shown, how such 3D pose tracking can help in problems like face recognition from videos.

References

1. Lanitis, A., Taylor, C., Cootes, T.: Automatic interpretation and coding of face images using flexible models. IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997) 743–756
2. Yuille, A.L., Cohen, D.S., Hallinan, P.W.: Feature extraction from faces using deformable templates. In: International Conference on Pattern Recognition. (1994)
3. Zhou, S., Chellappa, R., Moghaddam, B.: Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Trans. on Image Processing **11** (2004) 1434–1456
4. Hager, G.D., Belhumeur, P.N.: Efficient region tracking with parametric models of geometry and illumination. IEEE Trans. on Pattern Analysis and Machine Intelligence **20** (1998) 1025–1039
5. Jebara, T.S., Pentland, A.: Parameterized structure from motion for 3D adaptive feedback tracking of faces. In: IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR. (1997)
6. Pighin, F., Szeliski, R., Salesin, H.: Resynthesizing facial animation through 3D model-based tracking. In: Seventh International Conference on Computer Vision, Kerkyra, Greece. (1999) 143–150
7. Lu, L., Dai, X., Hager, G.: A particle filter without dynamics for robust 3D face tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, Washington, D.C. (2004)
8. Moon, H., Chellappa, R., Rosenfeld, A.: 3d object tracking using shape-encoded particle propagation. In: International Conference on Computer Vision. (2001)
9. Cascia, M.L., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000) 322–336
10. Birchfield, S.: An elliptical head tracker. In: Proceedings of the 31st Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California. (1997) 1710–1714
11. Fieguth, P., Terzopoulos, D.: Color-based tracking of heads and other mobile objects at video frame rates. In: IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR. (1997)
12. Doucet, A., Freitas, N.D., Gordon, N.: Sequential Monte Carlo methods in practice. Springer-Verlag, New York (2001)
13. Gordon, N.J., Salmond, D.J., Smith, A.F.M.: Novel approach to nonlinear/non-gaussian bayesian state estimation. In: IEE Proceedings on Radar and Signal Processing. Volume 140. (1993) 107–113
14. Jepson, A.D., Fleet, D.J., El-Maraghi, T.F.: Robust online appearance models for visual tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence **25** (2003) 1296–1311
15. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition. (2003)
16. Li, B., Chellappa, R.: Face verification through tracking facial features. Journal of the Optical Society of America A **18** (2001) 2969–2981