Computing 3-D Head Orientation from a Monocular Image Sequence

Thanarat Horprasert, Yaser Yacoob and Larry S. Davis
Computer Vision Laboratory
University of Maryland
College Park, MD 20742

Abstract

An approach for estimating 3D head orientation in a monocular image sequence is proposed. The approach employs recently developed image-based parameterized tracking for face and face features to locate the area in which a sub-pixel parameterized shape estimation of the eye's boundary is performed. This involves tracking of five points (four at the eye corners and the fifth is the tip of the nose). We describe an approach that relies on the coarse structure of the face to compute orientation relative to the camera plane. Our approach employs projective invariance of the cross-ratios of the eye corners and anthropometric statistics to estimate the head yaw, roll and pitch. Analytical and experimental results are reported.

1 Introduction

We present an algorithm for estimating the orientation of a human face from a single monocular image. The algorithm takes advantage of the geometric symmetries of typical faces to compute the yaw and roll components of orientation, and anthropometric modeling [3, 6] to estimate the pitch component. Estimating head orientation is central in vision-based animation, gaze estimation and as a component of inferring the intentions of agents from their actions. We seek an approach that requires no prior knowledge of the exact face structure of the individual being observed. The diversity of face and head appearances due to hair (head and facial) and eyeglasses in addition to articulated jaw motion and facial surface deformations leave very few features geometrically stable and predictable across individuals and head orientations. The nose is the only feature not subject to significant local deformations. In addition, the eyes are often visible (although occasionally covered by eye-glasses). For estimating head orientation, we assume that both eyes and the nose are visible, thus avoiding near-profile poses.

Several approaches have recently been proposed for estimating head orientation [5, 4]. In [5] the orientation is modeled as a linear combination of disparities between facial regions and several face models. Gee and Cipolla [4] estimate head orientation based on knowledge of the individual face geometry and assuming a *weak perspective* imaging model. Their method also depends on the distance between the eyes and mouth which often changes during facial expressions.

In this paper, a new approach for head orientation estimation is proposed. The approach employs recently developed image-based parameterized tracking [1] for face and face features to locate the area in which a sub-pixel parameterized shape estimation of the eye boundaries would be performed. This results in tracking of five points, four at the eye corners and the fifth at the tip of the nose. Although five points are not sufficient for recovering orientation in the absence of structure, we describe an algorithm that combines projective invariance of cross ratios from typical face symmetry and statistical modeling for face structure from anthropometry to estimate the three rotation angles. This approach consists of the following stages:

1. Region tracking of the face and the face features based on parameterized motion models (see [1]).
2. Sub-pixel estimation of eye-corners and nose-tip.
3. Computing 3D orientation from these five points.

of which stages (1) and (3) have been designed and implemented. In this paper we focus on the development and analysis of stage (3).

2 A Perspective Model for Head Orientation Recovery

In this section, we present the computational models for head orientation recovery based on projective invariants and anthropometric modeling of structure.

We employ a coordinate system fixed to the camera with the origin point being at its focal point (see Figure

The support of the Defense Advanced Research Projects Agency (DARPA Order No. C635) under Contract N00014-95-1-0521 is gratefully acknowledged.

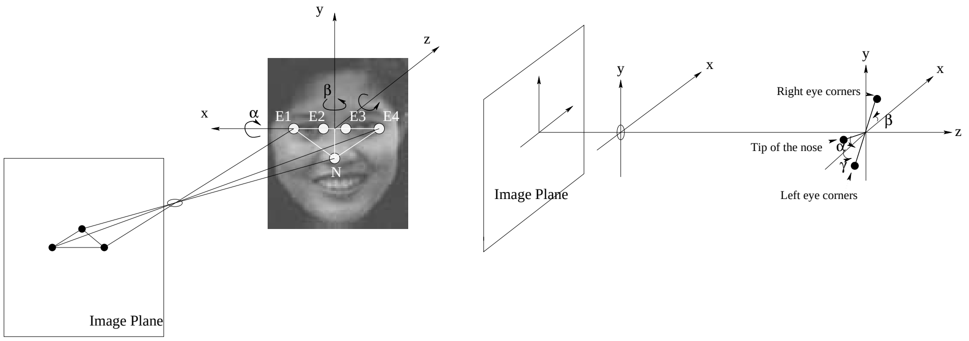

Figure 1: Geometric configuration of head orientation and coordinate system.

1). The image plane coincides with the XY-plane and the viewing direction coincides with the Z-axis. α, β, γ (pitch, yaw and roll, respectively) denote the three rotation angles of the head about the X,Y,and Z-axis, respectively.

Our model for head estimation assumes that the four eye-corners are co-linear in 3D; this assumption can be relaxed to account for a slight horizontal tilt in the eyes of Asian subjects (see [3] for statistics on the deviation of the eye corners from co-linearity).

Let upper case letters denote coordinates and distances in 3D and lower case letters denote their respective 2D projections. Let E_1, E_2, E_3 and E_4 denote the four eye corners and e_1, e_2, e_3 and e_4 denote their projection in the image plane. Let the coordinates of each point E_i be (X_i, Y_i, Z_i).

2.1 Roll Recovery

Roll is straightforward to recover from the image of the eye corner. From Figure 2 we see immediately that the head roll is

$$\gamma = \arctan \Delta y / \Delta x \tag{1}$$

where $\Delta y = {e_1}^y - {e_4}^y$ is the vertical distance and $\Delta x = {e_1}^x - {e_4}^x$.

2.2 Yaw Recovery

Let D denote the width of the eyes and D_1 denote half of the distance between the two inner eye corners (Figure 3). The head yaw is recovered based on the assumptions that (see Figure 3), if we assume that

1. $\overline{E_1E_2} = \overline{E_3E_4}$ (i.e., the eyes are of equal width).
2. E_1, E_2, E_3 and E_4 are collinear.

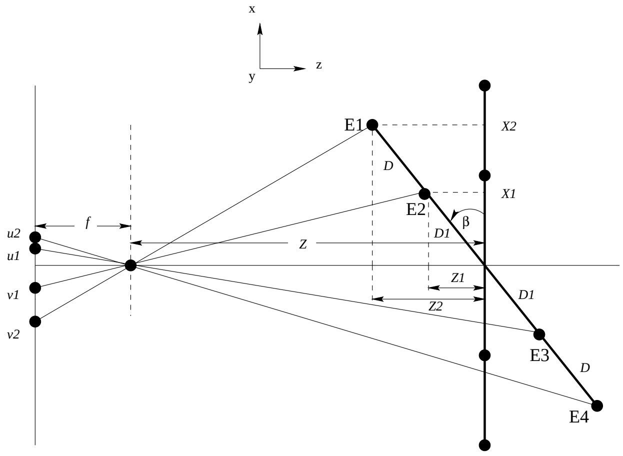

Figure 3: The perspective projection of eye corners while the head is rotating about Y-axis.

Then from the well-known projective invariance of the cross-ratios we have

$$I_1 = \frac{(u_2 - u_1)(v_1 - v_2)}{(u_2 - v_1)(u_1 - v_2)} = \frac{D^2}{(2D_1 + D)^2} \tag{2}$$

which yields

$$D_1 = \frac{DQ}{2} \tag{3}$$

where

$$Q = \frac{1}{\sqrt{I_1}} - 1$$

From perspective projection we obtain

$$u_1 = \frac{fX_1}{Z + Z_1} = \frac{fD_1 \cos\beta}{Z + D_1 \sin\beta} \tag{4}$$

$$u_2 = \frac{fX_2}{Z + Z_2} = \frac{f(D + D_1)\cos\beta}{Z + (D + D_1)\sin\beta} \tag{5}$$

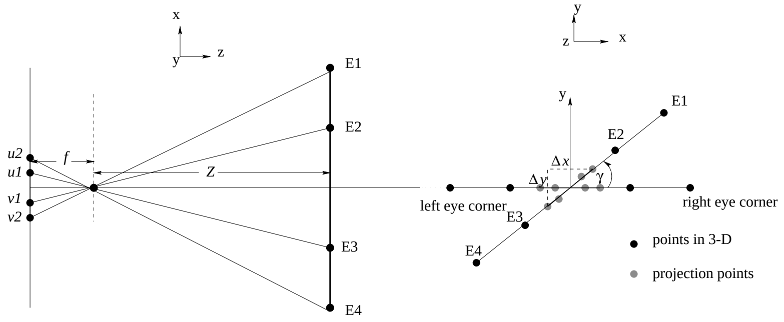

Figure 2: The perspective projection of eye corners while the head rotates about Z-axis.

$$v_1 = \frac{fX_1}{Z - Z_1} = \frac{fD_1 \cos\beta}{Z - D_1 \sin\beta} \quad (6)$$

$$v_2 = \frac{fX_2}{Z - Z_2} = \frac{f(D + D_1)\cos\beta}{Z - (D + D_1)\sin\beta} \quad (7)$$

where f is the focal length of the camera. From (4), we obtain

$$Z = \frac{fX_1}{u_1} + Z_1 = Z_1(\frac{fX_1}{u_1 Z_1} - 1) = Z_1 S \quad (8)$$

where

$$S = \frac{f}{u_1 tan\beta} - 1 \quad (9)$$

From (9), we can determine the head yaw (β)

$$\beta = \arctan\frac{f}{(S-1)u_1} \quad (10)$$

However, since u_1 is measured relative to the projection of the midpoint of $\overline{E_1 E_4}$ (which is unknown) we need to determine S and u_1 from the relative distances among the projections of four eye corners.

From (2) - (8), we obtain a quadratic equation in S:

$$\frac{\Delta u}{\Delta v} = \frac{u_2 - u_1}{v_1 - v_2} = -\frac{(S-1)(S-(1+(2/Q)))}{(S+1)(S+(1+(2/Q)))} \quad (11)$$

So,

$$S = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \quad (12)$$

where

$$\begin{aligned} A &= (\Delta u/\Delta v + 1) \\ B &= ((2/Q) + 2)(\Delta u/\Delta v - 1) \\ C &= ((2/Q) + 1)(\Delta u/\Delta v + 1) \end{aligned}$$

To determine u_1, we employ another two cross-ratio invariants

$$\frac{(u_2 - u_1)v_1}{(u_1 - v_1)u_2} = -\frac{DD_1}{2D_1(D + D_1)} = -\frac{1}{2+Q} = M \quad (13)$$

$$\frac{(v_1 - v_2)u_1}{(u_1 - v_1)v_2} = -\frac{DD_1}{2D_1(D + D_1)} = M \quad (14)$$

From (13) and (14) it can be shown that

$$\frac{[\Delta v \Delta u M(u_1 - v_1)] - [M^2(u_2 - v_2)(u_1 - v_1)^2]}{\Delta v(M(u_1 - v_1) - \Delta u)} = u_1 \quad (15)$$

By replacing (12) and (15) in (10), we can now determine the head yaw angle (β). Note that β depends only on the *relative* distances among four corners of the eyes and the focal length while being independent of the face structure and the distance of the face from the camera. It is also not influenced by other parameters such as the translation of the face along any axis.

2.3 Pitch Recovery

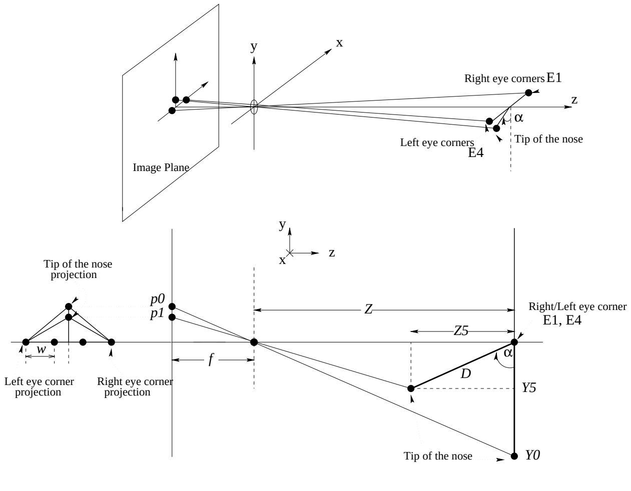
y
x
Right eye corners E1
z
α
Tip of the nose
Left eye corners
E4
Image Plane
y
x
z
Tip of the nose projection
p0
p1
Z
Right/Left eye corner E1, E4
Z5
w
f
α
D
Y5
Left eye corner projection
Right eye corner projection
Tip of the nose
Y0

Figure 4: *Top*: The eye corners and nose tip positions and the 3D imaging model for pitch (rotation around X) *Bottom*: Perspective model for pitch recovery.

Let D_2 denote the 3D distance between the outer corner of each eye and the tip of the nose (for a symmetric face), p_0 denote the projected length of the bridge of the nose when it is parallel to the image plane and p_1 denote the observed length of the bridge of the nose at the unknown pitch. Let (X_0, Y_0, Z_0) and (X_5, Y_5, Z_5) denote the 3D coordinates of the tip of the nose at 0 degrees and at the current angle α. From the perspective projection (Figure 4), we obtain

$$\frac{f}{Z} = \frac{p_0}{Y_0} = \frac{p_0}{D_2} \Longrightarrow Z = \frac{fD_2}{p_0} \qquad (16)$$

$$\frac{f}{Z - Z_5} = \frac{p_1}{Y_5} = \frac{p_1}{D_2 \cos\alpha} \Longrightarrow Z = \frac{fD_2 \cos\alpha}{p_1} + D_2 sin\alpha \qquad (17)$$

From (16) and (17) it can be shown that

$$(p_1^2 + f^2)\sin^2\alpha - \frac{2fp_1^2}{p_0}\sin\alpha + f^2(p_1^2/p_0^2 - 1) \quad = \quad 0$$

The estimated pitch angle, α, can be computed by :

$$\alpha = \arcsin[E] \qquad (18)$$

where

$$E \quad = \quad \frac{f}{p_0(p_1^2 + f^2)}[p_1^2 \pm \sqrt{p_0^2 p_1^2 - f^2 p_1^2 + f^2 p_0^2}\ ]$$

Computing E requires estimating p_0 which is not generally known. Instead, we obtain it by first categorizing the observed face with respect to the variables of gender, race and age (see [2]) and then use tabulated anthropometric data to estimate the mean and expected error (used to estimate accuracy of pitch recovery see Section 3) of p_0. Let N denote the average length of the nasal bridge and E denote the average length of the eye fissure (Biocular width).

By employing these statistical estimates for the face structure variables, p_0 can be estimated :

$$\begin{aligned} p_0 &= \frac{fN}{z} \\ w &= \frac{fE}{z} \Longrightarrow z = \frac{fE}{w} \\ p_0 &= \frac{Nw}{E} \end{aligned}$$

where w is the length of projective eye fissure in the image plane.

3 Error Analysis Simulation

In this section the effects of both image error (localization of eye corners) and model errors (due to expected variations in length from the anthropometric models) on orientation recovery are presented. In this analysis, we assume that the structure is that of an average adult American male. Thus, the expected length of the nasal bridge is 50mm, and the expected length of eye fissure is 31.3mm. The distance between model and camera is 500mm, and the focal length is 16mm. We also assume a pixel size of 16 μm. We first explore the sensitivity of the image features upon which the recovery of orientation is based to changes in the orientation as a function of orientation. Figure 6 predicts that yaw and roll will be least accurately estimated around 0 degrees while pitch is least accurately recovered at an angle ϕ slightly offset from 0 (see explanation below).

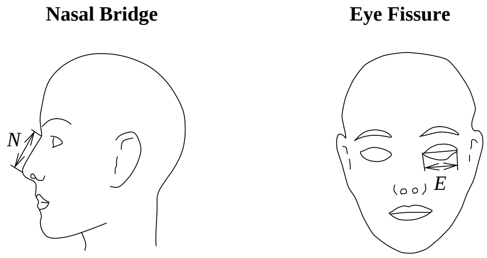

Figure 5: Examples of face anthropometry used in our computation; length of nasal bridge and length of eye fissure.

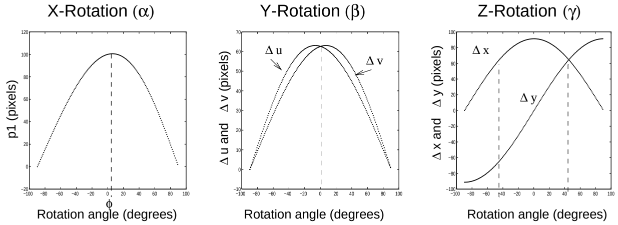

Figure 6: The sensitivity of critical image distances to the head orientation, u_1 for α, Δu and Δv for β, and dx and dy for γ

Figure 7 illustrates the localization accuracy of image measurements required to obtain a given accuracy (0.5 to 5.0 degrees) in recovering orientation as a function of absolute orientation. As can be observed very high localization accuracy is needed for yaw and roll of about 0 degrees. We call these angles *degenerate angle*. While the degenerate angles for yaw and roll are independent of any length scales, the degenerate angle for pitch depends on the angle $\phi = \arcsin(N/Z)$ where N is the length of nasal bridge and Z is the distance between the face and the camera. In this particular analysis, $\phi \sim 5.74$ degree $(arcsin\frac{50}{500})$. Generally, this angle is the angle for which the line of sight to the nose tip is tangent to the circle of radius N centered on the

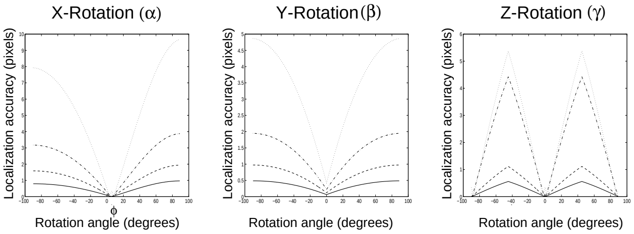

Figure 7: The localization accuracy of the rotation about each axis, while maintaining 0.5(solid line), 1(dash line), 2(dash+point line), and 5(dot line) degrees error.

face (see Figure 8).

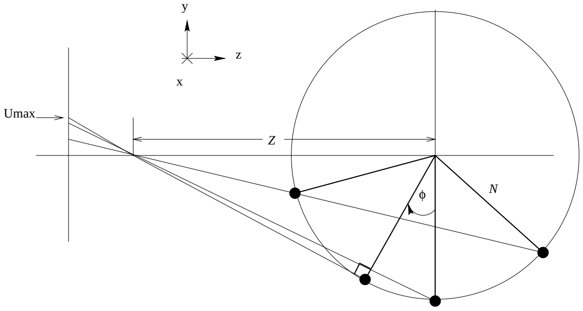

Figure 8: An illustration showing characteristic of rotating about X-axis. By considering rotating from -90 to 90 degree, the projective point on image plane will reach the highest point when the angle of rotation is ϕ degree. This tangent angle is enfluenced by the distance between the origin point of rotation and the len, and the radius of the rotation.

In addition to depending on the localization accuracy of the image processing, the recovery of pitch also depends on the deviation of the observed face from the assumed face anthropometry as illustrated in Figure 9.

The horizontal axis of these graphs represents the actual pitch while the vertical axis is the estimated pitch employing an average face structure (i.e., 50mm for nasal bridge and 31.3mm for eye width); four cases are shown

(a) The actual length of the nasal bridge of the model varies from 45-50mm.

(b) The actual length of the nasal bridge of the model varies from 50-55mm.

(c) The actual eye size of the model varies from 29.8-31.3mm.

(d) The actual eye size of the model varies from 31.3-32.8mm.

As we can see from the graphs, the error is highest when the rotation is close to the degenerate angle ϕ.

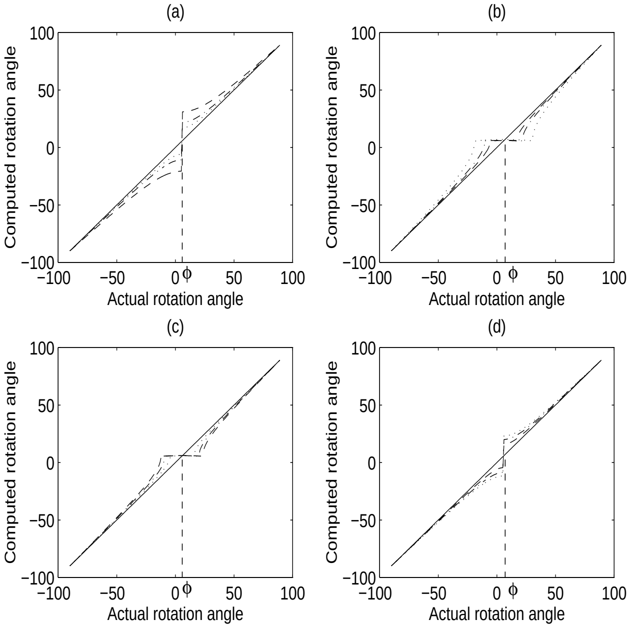

Figure 9: An illustration showing errors of applying our computation models to various faces structures.

4 Experimental Results

In this section we provide some experimental results on real image sequences captured at 25 frames per second. Some frames of the sequence are shown in Figure 10 and Figure 11. In Figure 10, the plots of α, β, and γ are shown. The five feature points were selected by hand for these examples.

Another experiment was performed to compare pitch recovery using anthropometric models and an individual's actual face structure. For this purpose, two image sequences of a Caucasian male and an Asian female were captured. The results are shown in Figure 11. The plots show the differences between using individual structure (true 3D measurement of the subject's features) and the anthropometry in the computational models. The means of the face structures of adult American Caucasian male and female were employed in this experiment. In both cases, the pitch is recovered with an error predicted by our model.

5 Summary

We have presented an approach for computation of head orientation by employing projective invariants and statistical analysis of face structure. We divided the computation into separate estimation of the orientation about Z, Y, and X-axis, respectively. The rotations of the head about Z and Y-axis can be computed

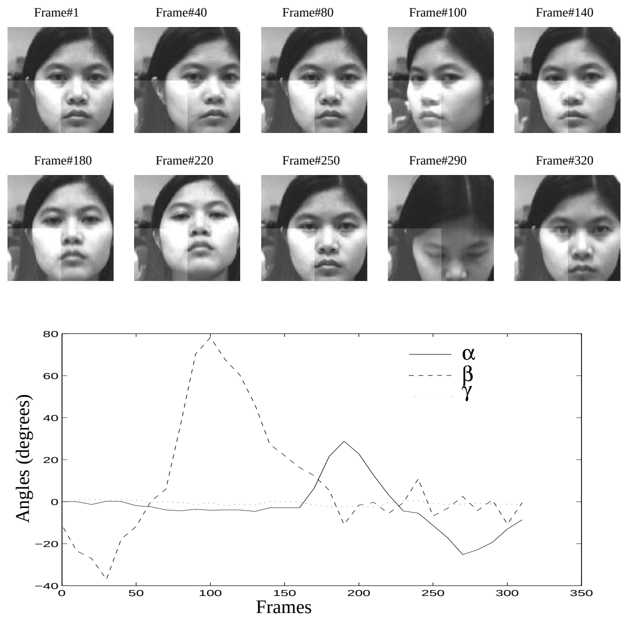

Figure 10: Some images in a sequence, and the result of pose estimation. The plots shows the sequences of rotation angles through the image sequence.

directly from the relative distances among corners of the eyes and the camera focal length. The estimation of the orientation of the head about X-axis employed anthropometric analysis of face structure. A preliminary set of experiments on real images has demonstrated the performance of our proposed approach.

References

[1] M. J. Black and Y. Yacoob, "Tracking and Recognizing Facial Expressions in Image Sequences, using Local Parameterized Models of Image Motion", *ICCV,* 1995, 374-381.

[2] R. Chellappa, C.L. Wilson and S.A. Sirohey. Human and machine recognition of faces: A survey. *Proc. of IEEE,* Vol 83, 1995, pp.705–740.

[3] L.G. Farkas, "Anthropometry of the Head and Face" 2nd edition, *Raven Press*, 1994.

[4] A. Gee and R.Cipolla, "Estimating Gaze from a Single View of a Face," *ICPR'94*, 758-760, 1994.

[5] A. Tsukamoto, C. Lee and S. Tsuji, "Detection and Pose Estimation of Human Face with Synthesized Image Models," *ICPR'94*, 754-757, 1994.

[6] J. Young, *Head and Face Anthropometry of Adult U.S. Citizens*, Government Report DOT/FAA/AM-93/10, July 1993.

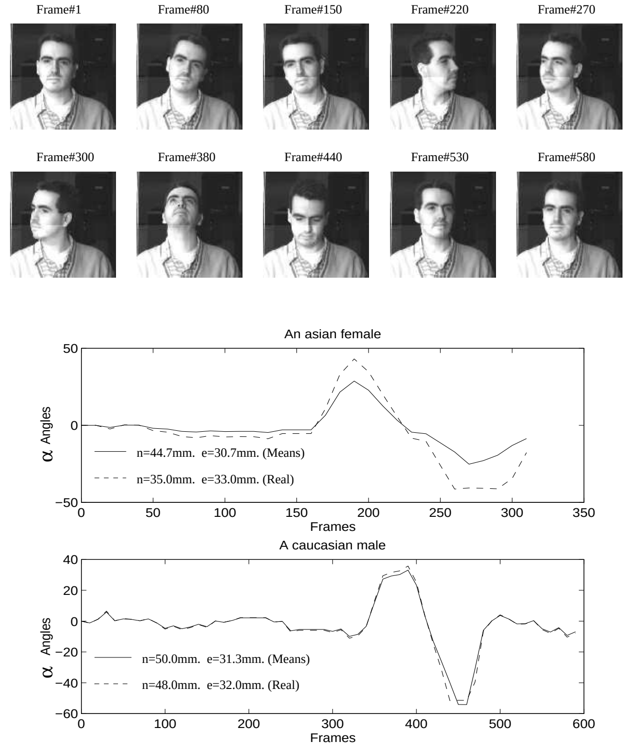

Figure 11: The error when applying our approach to different face structures. (*solid line*) is the plot of α computed by employing statistic anthropometry, while (*dash line*) is the plot of α computed by utilizing the real measurements of model's face structure.