REAL-TIME HEAD POSE ESTIMATION USING RECONSTRUCTED 3D FACE DATA FROM STEREO IMAGE PAIR

Sabri Gurbuz[†‡] *and Naomi Inoue*[†‡]

[†]*NICT Cognitive Information Science Laboratories*
[‡]*ATR Cognitive Information Science Laboratories, Kyoto, Japan*
{*sabrig, na-inoue*}*@atr.jp*

ABSTRACT

In this paper, we describe a real-time 3D human head pose estimation algorithm by means of eye tracking and on-line reconstructed 3D face data from a stereo image pair. A frame by frame based face coordinate system is formed using eye locations and a face plane which doesn't rely on distinctive facial features. The rest of the algorithm follows the conventional approach of pose estimation between the face coordinate system and a reference coordinate system without any initialization and model fitting. By not relying on arbitrary data on the face such as nose ridge line or a mouth corner points, we minimize the impact of errors which may come from their 3D estimates and facial expression changes. Experimental results demonstrate the accuracy and robustness of the algorithm.

Index Terms— Stereo eye tracking, 3D face reconstruction, 3D head pose estimation.

1. INTRODUCTION

We present an algorithm for estimating the 3D head pose of a person from reconstructed 3D face data. The algorithm takes advantage of a fast 2D based eye tracking method to reconstruct 3D face from stereo image pair to compute roll, pitch, yaw components of the head pose and the head translation at each frame. Determining the head pose is one of the central problems in vision based human-computer interface applications. In general, a person's head pose is largely related to person's intention and attention. Therefore, capturing and understanding human head motion has become one of the active research areas in computer vision recently [1, 2, 3]. Hence, we seek an approach that requires no prior knowledge of the face structure of the individual(s) being observed. The eyes, which are salient even when eyeglasses are present, have the consistent gradient feature with respect to the chins and nose bridge. For estimating head pose, as an initial step, an eye tracking algorithm quickly screens the 2D image frame for detecting face candidates for the purpose of further processing.

There are various commercial products available for the head pose estimation. However, they require special sensors or markers placed on the user's face or head. For example, the eye gaze tracking based systems employing infrared illumination guarantee reliable detection of eye locations, but fails under direct sunlight where such system wouldn't be suitable for head pose estimation in general such as public domain face recognition. Most approaches in the literature employ geometric feature based, model based, or learning based algorithms either using 2D image data [4, 5, 6, 7, 8, 9] or range/stereo data [1, 3, 2, 10, 11, 12, 13, 14] for the head pose estimation.

In this paper, we have extended previous efforts by proposing an eye tracking based horopter estimation for 3D face reconstruction of the user(s). The stereo vision based 3D reconstruction and head pose estimation, which was overlooked by many researchers in the past, provides an attractive method. However, it involves several challenges such as real-time 3D face reconstruction and selection of facial data for the head pose. Our head pose estimation approach consist of four steps; in the first step, face candidate regions are located in the left camera image by an eye pattern search algorithm [15], which searches the pattern between the eyes. During step two, the resulting face candidate regions are further verified by the stereo epipolar constraint. Step three utilizes the horopter information of the face for 3D reconstruction of face information using the stereo image pair. Finally, in step four, a face plane is estimated from 3D reconstructed face data. Then, a face coordinate system is formed from 3D eye locations and the face plane for the real-time 3D head pose estimation.

This work is organized as follows. In section 2, tracking the pattern between the eyes and its stereo extension are described briefly. Section 3 discusses the reconstruction of 3D face data of the individual(s) being observed on the fly. Section 4 describes how to form a frame by frame based 3D face coordinate system. Then, the 3D head pose estimation algorithm is described in section 5. Experimental setup and discussions are presented in Section 6. Finally, conclusion is given in Section 7.

2. TRACKING THE PATTERN BETWEEN-THE-EYES

Our 3D head pose estimation algorithm doesn't rely on a specific eye tracker. Hence, any eye tracking algorithm would be suitable for the work. In this implementation, the pattern between the eyes (eye locations have lower intensity than the cheek and nose ridge) are detected and tracked with size updated pattern matching. To cope with scales of faces, various scales of patterns are considered during the detection, and an appropriate scale is selected accordingly for the tracking. The algorithm calculates the intermediate representation of the input image called "Integral image", described in [16]. Then, a six segmented rectangular (SSR) filter is used for fast filtering of bright-dark relations of the eye region in the image. For the initial verification of the candidates, a support vector machine (SVM) algorithm is employed (interested readers may refer to Ref. [15] for the details of the algorithm). Using the epipolar constraint, the corresponding between-the-eye pattern is searched along the epipolar line in the right camera image using a correlation based template matching algorithm for estimating the stereo eye locations and the user's head horopter.

3. RECONSTRUCTION OF 3D FACE DATA

Stereo computer vision algorithms use disparity along with camera calibration parameters for the computation of 3D (X,Y,Z) coordi-

Fig. 1. Screen capture of the left and right camera images, and OpenGL plot of texture mapped 3D point clouds reconstructed by the stereo algorithm.

nates of an object point. Disparity algorithms often rely on search mechanisms where a template window in the left camera image is compared with the windows on the epipolar line in the right camera image using various measures such as squared difference or normalized correlation coefficient methods. That is, the matching algorithm needs to calculate a similarity measure for a 2D template window A in the left image to a 2D window B in the right image. We can express A and B in general as in the following.

$$A = \gamma\tilde{A} + \bar{A} \quad (1)$$

$$B = \Gamma\tilde{B} + \bar{B} \quad (2)$$

where γ and Γ represent arbitrary gain/scaling values of the left and right cameras, respectively. $\bar{X}$ stands for the average value and $\tilde{X}$ represents the intrinsic shape or texture properties of the window data. That is, under the ideal conditions $\tilde{A}$ and $\tilde{B}$ are identical if they belong to same object patch. However, in real world conditions, disparity search algorithms may fail because of video noise or lack of unique intrinsic shape properties where $\tilde{A}$ and $\tilde{B}$ carry no information. The video noise may occur due to differing lens focuses, differing viewing angles, and uneven lighting effects on the left and right cameras.

In order to reduce the number of failures or false matches, the horopter of the face is estimated from stereo eye tracking on the fly. Then, the search domain of the disparity algorithm is restricted for the face. Resulting disparities are further verified by the consistency check within the pixel neighborhoods. In this implementation, missing disparities are estimated by linear interpolation of their neighbors. Fig. 1 shows the tracked eye locations in the left and right camera images, and the reconstructed 3D face data from the stereo pair.

4. FORMING OF 3D FACE COORDINATE SYSTEM

A vector on a plane and the plane's normal can describe an arbitrary 3D coordinate system. Hence, having 3D eye locations and reconstructed 3D face data, our goal is to form a frame by frame based 3D face coordinate system for the head pose estimation.

Formation of a face coordinate system based on 3D locations of facial features such as nose ridge line, nose tip or mouth corners seems a reasonable approach, however additional to their difficulty of tracking, these features are either deforming during natural speech or their sizes are person dependent. Therefore, we propose to utilize 3D eye locations and a face plane that doesn't rely on distinctive facial futures.

The concept of the proposed 3D face coordinate system is as follows; first, the face plane is estimated from the 3D face data between the eyes and chin, but excluding the strip that contains the mouth and nose. However, during extreme head orientations 3D face data may contain outliers. Therefore, an iterative least square method is utilized to eliminate the points with certain high errors using the histogram of errors between the face data and the estimated plane. Iteration stops after 2 or 3 iterations and doesn't require heavy computation. The face plane parameters ($aX + bY + cZ + d = 0$) is estimated by using the least square solution with the remaining 3D face data. Thus, the normal of the face plane is available and can be described by the vector $\overrightarrow{V}_z = (a, b, c)^t$.

Let $\mathbf{E_1}$ and $\mathbf{E_2}$ be the positions of the left and right eyes on the face plane in 3D space such that $\mathbf{E_i} = (X_i, Y_i, \hat{Z}_i)^T$ where $\hat{Z}_i$ is the re-calculated Z-value on the face plane given its X_i and Y_i values. Now, an equation for the eye line in 3D space ($\mathbf{E_{line}}$) can be defined by utilizing the points $\mathbf{E_1}$ and $\mathbf{E_2}$ as,

$$\mathbf{E_{line}} = \mathbf{E_1} + t * \overrightarrow{V}_x, \quad (3)$$

Where t is a scalar value ($t \in R$) for a point on $\mathbf{E_{line}}$, and $\overrightarrow{V}_x = \overrightarrow{\mathbf{E_1}\mathbf{E_2}}$ which defines a vector perpendicular to the normal of the face plane.

The cross product of the vector $\overrightarrow{V}_z$, which is the normal of the face plane, and the vector $\overrightarrow{V}_x$, which is obtained from the 3D eye line in Eqn. 3, produces the vector $\overrightarrow{V}_y$ which is perpendicular to both $\overrightarrow{V}_x$ and $\overrightarrow{V}_z$. These three vectors form the face coordinate system. Hence, x-axis and z-axis of the face coordinate system are locked to 3D eye locations and the 3D face plane, respectively. Therefore, the forming of the 3D face coordinate system is repeatable frame by frame basis. The following section describes the transformation, which represent the 3D head pose, between the face coordinate system and the global (reference) coordinate system.

5. 3D HEAD POSE ESTIMATION

A rotation only transformation between two coordinate systems can be expressed by

$$\left[\begin{array}{ccc} \overrightarrow{x'} & \overrightarrow{y'} & \overrightarrow{z'} \end{array} \right] = R(\phi) \left[\begin{array}{ccc} \overrightarrow{x} & \overrightarrow{y} & \overrightarrow{z} \end{array} \right] \quad (4)$$

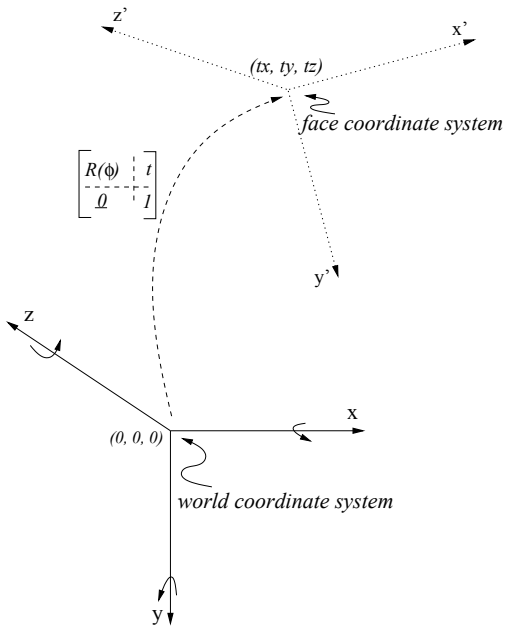

Fig. 2. General transformation between the world coordinate system and the face coordinate system, where $\underline{0} = [0\ 0\ 0]$, and $\underline{t} = (t_x,\ t_y,\ t_z)^T$.

where $\overrightarrow{x'}$, $\overrightarrow{y'}$ and $\overrightarrow{z'}$ are resulting unit length vectors after rotating the standard unit vectors ($\overrightarrow{x}$, $\overrightarrow{y}$, and $\overrightarrow{z}$) around the origin of the world coordinate system by $R(\phi)$.

In general, a face coordinate system can be thought of a rotated and translated form of a world coordinate system. That is, the face coordinate system and the world coordinate system are related by a 3×3 rotation matrix $R(\phi)$ and a 3×1 translation vector $\underline{t}$ as shown in Figure 2. Independent from the translation $\underline{t}$, the vectors $\overrightarrow{x'}$, $\overrightarrow{y'}$ and $\overrightarrow{z'}$ define the axis directions of the face coordinate system which are defined in this implementation by

$$\overrightarrow{x'} = \overrightarrow{V}_x / \parallel \overrightarrow{V}_x \parallel, \quad (5)$$

$$\overrightarrow{y'} = \overrightarrow{V}_y / \parallel \overrightarrow{V}_y \parallel, \quad (6)$$

$$\overrightarrow{z'} = \overrightarrow{V}_z / \parallel \overrightarrow{V}_z \parallel . \quad (7)$$

Thus, solution to the 3D head pose matrix $R(\phi)$ in Equation 4 becomes trivial and can be expressed by [17]

$$R(\phi) = R_z(\varphi)R_y(\vartheta)R_x(\psi) = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}. \quad (8)$$

where ψ, ϑ, and φ are rotation angles around x-, y-, and z-axises, respectively. The rotation angles around their respective axises that satisfy Equation 8 are given by

$$\varphi = atan(r_{21}/r_{11}), \quad (9)$$

$$\vartheta = atan(-r_{31}/\sqrt{r_{32}^2 + r_{33}^2}), \quad (10)$$

$$\psi = atan(r_{32}/r_{33}). \quad (11)$$

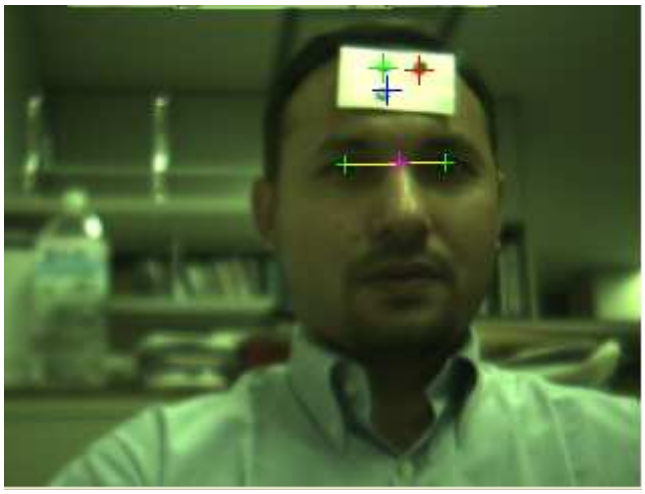

Fig. 3. Marker object on the forehead for the marker based pose estimation.

6. EXPERIMENTAL SETUP AND DISCUSSION

Videre stereo vision hardware and the SVS software are utilized for the implementation. Calibration and rectification of cameras are done automatically using the SVS library. The SVS software can capture stereo video sequences and reconstructs 3D data of the stereo pair at 30Hz with full image resolution of 320×240. However, for our goal, the 3D reconstruction region of interest (ROI) is the user's face area, so we restrict the disparity search region around the horopter of the face. Therefore, the reconstructed 3D data outside the face with a different depth has an inaccurate 3D estimate as can be seen in Fig. 1. The 3D coordinate values are calculated with respect to a world coordinate system. The world coordinate system (origin) in this implementation is defined to be the focal point of the left camera and is a right-handed coordinate system.

To access the accuracy of the head pose estimation algorithm, comparison with measurements obtained from a marker based head pose estimation was performed. Three circular black markers with a 5 millimeter radius are placed 25 millimeters apart forming a 90 degree clockwise rotated L shape on the user's forehead as seen in Fig. 3. A stereo processing algorithm is utilized to estimate the 3D coordinates of the marker positions. The accuracy of the marker based head pose estimation algorithm is ± 3 degrees due to jitters in the marker position detection.

In Fig. 4, the estimated angles from the proposed algorithm are compared with the values obtained by the marker based algorithm. Naturally, the forehead, where the markers were placed, has different pose than the face which is reflected in the plots by their respective offset values. For the x-, y- and z-axis angle data depicted in Fig. 4, correlation coefficients are 0.87, 0.92, and 0.98 respectively. The correlation coefficient for the rotations around the x-axis is lower compared to the results of y- and z-axises due to jitters in eye locations detected by the eye tracking algorithm. However, the jitter in eye tracking can be corrected from the 3D face structure information or using a better eye tracker.

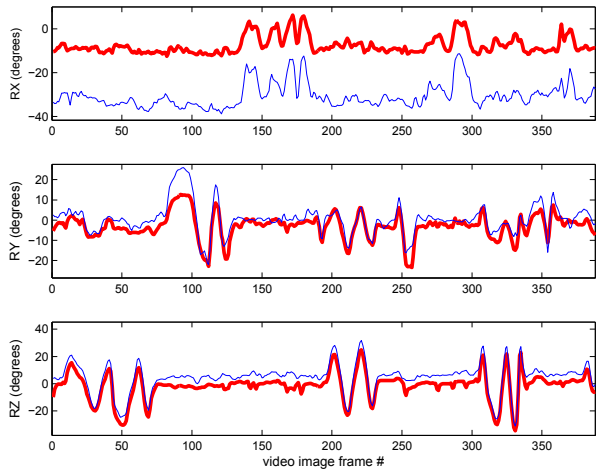

Fig. 4. Comparison of pose angles between the results given by the proposed algorithm (thick red lines) and with those measured by marker based algorithm (thin blue lines).

7. CONCLUSIONS

We have described a robust 3D head pose estimation algorithm which is suitable for human-computer interface applications in real world conditions. The proposed algorithm is model free and initialization free and can estimate 3D head pose information from a single image pair. It is also robust against facial expressions and person dependent facial features such as nose shape.

Future work includes extension to a multiple person head pose estimation.

Acknowledgment

Thanks are due to Shinjiro Kawato for the original eye tracking algorithm extended in this paper.

8. REFERENCES

[1] S. Malassiotis and M. G. Strintzis, "Robust real-time 3d head pose estimation from range data," *Pattern Recognition*, vol. 38, pp. 1153–1165, 2005.

[2] Y. Matsumoto and A. Zelinsky, "An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[3] E. Seemann, K. Nickel, and R. Stiefelhagen, "Head pose estimation using stereo vision for human-robot interaction," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[4] T. Horprasert, Y. Yacoob, and L.S. Davis, "Computing 3-d head orientation from a monocular image sequence," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996.*, 1996.

[5] M. Malciu and F. Preteux, "A robust model-based approach for 3d head tracking in video sequences," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[6] M.D. Cordea, E.M. Petriu, N.D. Georganos, D.C. Petriu, and T.E. Whalen, "Real-time 2(1/2)-d head pose recovery for model-based video-coding," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, pp. 1007–1013, 2001.

[7] Q. Chen, H. Wu, T. Fukumoto, and M. Yachida, "3d head pose estimation without feature tracking," in *3rd. International Conference on Face and Gesture Recognition*, 1998.

[8] R. Lopez and T.S. Huang, "3d head pose computation from 2d images: templates versus features," in *Proceedings of the 1995 International Conference on Image Processing (Vol.2)*, 1995.

[9] C.-P. Lu, G.D. Hager, E. Mjolsness, and C.A. Sunnyvale, "Fast and globally convergent pose estimation from video images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 610–622, 2000.

[10] R. Yang and Z. Zhang, "Model-based head pose tracking with stereovision," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[11] J.-G Wang, E.T. Lim, R. Venkateswarlu, and E. Sung, "Stereo head/face tracking and pose estimation," in *Seventh International Conference on Control, Automation, Robotics And Vision (ICARCV'02)*, 2002.

[12] M. Harville, A. Rahimi, T. Darrell, G. Gordon, and J. Woodfill, "3d pose tracking with linear depth and brightness constraints," in *Seventh International Conference on Computer Vision (ICCV'99) - Volume 1*, 1999.

[13] M. Xu and T. Akatsuka, "Detecting head pose from stereo image sequence for active facerecognition," in *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, 1998.

[14] H. Song and K. Sohn, "3d head pose estimation using range images for face recognition," in *International Conference on Control, Automation, Robotics And Vision (ICARCV'04)*, 2004.

[15] S. Kawato and N. Tetsutani, "Scale adaptive face detection and tracking in real time with ssr filter and support vector machine," in *Proc. of ACCV, vol. 1*, 2004.

[16] P. Viola and M. Jones, "Robust real-time object detection," in *Second International Workshop on Statistical and Computational Theories of Vision- Modeling, Learning, Computing, and Sampling, Vancouver, Canada,*, 2001.

[17] L. Sciavicco and B. Sicilliano, *Modelling and Control of Robot Manipulators*, Springer, 2001.