2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

# Robust Unified Stereo-Based 3D Head Tracking and Its Application to Face Recognition

Kwang Ho An and Myung Jin Chung, *Senior Member, IEEE*

**Abstract—This paper investigates the estimation of 3D head poses and its identity authentication with a partial ellipsoid model. To cope with large out-of-plane rotations and translation in-depth, we extend conventional head tracking with a single camera to a stereo-based framework. To achieve more robust motion estimation even under time-varying lighting conditions, we incorporate illumination correction into the aforementioned framework. We approximate the face image variations due to illumination changes as a linear combination of illumination bases. Also, by computing the illumination bases online from the registered face images, after estimating the 3D head poses, user-specific illumination bases can be obtained, and therefore illumination-robust tracking without a prior learning process can be possible. Furthermore, our unified stereo-based tracking is approximated as a linear least-squares problem; a closed-form solution is then provided. After recovering the full-motions of the head, we can register face images with pose variations into stabilized-view images, which are suitable for pose-robust face recognition. To verify the feasibility and applicability of our approach, we performed extensive experiments with three sets of challenging image sequences.**

## I. Introduction

An accurate estimation of 3D head position and orientation is important in many applications. 3D head pose information can be used in human-computer interfaces (HCI), active telecommunication, virtual reality, and visual surveillance. In addition, a face image aligned in terms of the recovered head motion would facilitate face recognition and facial expression analysis. Thus, many approaches to recover 3D head motion have been proposed [1]-[4]. One is to use distinct image features. This approach works well when the features may be reliably tracked over the image sequence. When this is not possible, using a 3D head model to track the entire head region is more reliable. There have been several model-based techniques to track a human head in 3D space.

Cascia et al. [2] developed a fast 3D head tracker that models a head as a texture-mapped cylinder. The head pose of the input image is treated as a linear combination of a set of 24 warping templates (4 templates × 6 motion parameters) and a set of 10 illumination templates that are obtained through a prior learning process. While simple and effective, use of a small number of static templates appears unable to cope with fast and large out-of-plane rotations and translation in-depth.

Xiao et al. [3] presented a method to recover the full-motion (3 rotations and 3 translations) of the head using a cylindrical model. They used the iteratively re-weighted least squares technique to deal with non-rigid motion and occlusion. For tracking, the templates are dynamically updated to diminish the effects of self-occlusion and gradual lighting changes. However, since their method is not considering illumination correction explicitly, their tracker is not likely to work well under time-varying illumination conditions.

The above two methods model a human head as a 3D cylinder. However, since the human head is not a 3D cylinder, modeling inaccuracies between the actual and approximated head modes can be significant. This inherent modeling error may degrade the accuracy in motion estimation.

Blanz and Vetter [4] proposed an algorithm to fit 2D face images with 3D Morphable Models to estimate the head pose. Although the head pose can be estimated accurately, their method suffers from the cost of 3D data acquisition and processing. The average processing time for each frame is around 30 seconds. This is too slow for real-time applications.

All the methods described above are based on head pose estimation using only a single camera. Generally, 3D head tracking with a single camera is not robust to fast and large our-of-plane rotations and translation in-depth.

With consideration of all of these issues, the coverage of this paper is as follows. As in [5], we model the shape of a human head as a partial 3D ellipsoid-a reasonable approximation to the actual head. Also, to complement the weakness of a single camera system, we extend conventional head tracking with a single camera to a stereo-based framework. Through the use of the extra information obtained from stereo images, coping with large out-of-plane rotations and translation in-depth is now tractable (or at least easier than with a single camera). Furthermore, we incorporate illumination correction into this stereo-based framework to allow for more robust motion estimation even under time-varying illumination conditions. We approximate the face image variations due to illumination changes as a linear combination of illumination bases. By computing the illumination bases online from the registered face images, after estimating the 3D head poses, user-specific illumination bases can be obtained, and therefore illumination-robust tracking without a prior learning process can be possible.

To verify the applicability of the proposed approach, we apply our head tracker to face recognition. Generally, the

Manuscript received September 15, 2009.

K. H. An and M. J. Chung are with the Electrical Engineering and Computer Science Department, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea (e-mail: akh@cheonji.kaist.ac.kr, mjchung@ee.kaist.ac.kr).

978-1-4244-5040-4/10/$26.00 ©2010 IEEE

performance of face recognition deteriorate with changes in pose, illumination, and other disturbing factors, among which pose variation is the most difficult one to deal with [6]. Therefore, face registration is the key of robust face recognition. After recovering the full-motions of the head by the proposed head tracker, face images with pose variations can be registered into stabilized-view images, which are suitable for pose-robust face recognition.

The remainder of the paper is organized as follows. Section II presents a unified 3D head pose estimation method including online illumination correction. Section III explains how to generate stabilized and mirrored texture maps, which are suitable for frontal face recognition, by using the unified stereo-based tracking framework proposed in Section II. In Section IV, we provide extensive experimental results with three sets of challenging image sequences. Section V presents conclusions and discussions.

## II. Unified Stereo-Based 3D Head Pose Estimation

Generally, image-based tracking is based on the brightness change constraint equation (BCCE). The BCCE for image velocity estimation arises from the assumption that image intensity does not change from one frame to the next. However, this assumption does not hold true under real-world conditions. Tracking based on the minimization of the sum of squared differences between the input and reference images is inherently susceptible to changes in illumination. Hence, we need to consider the effect of ambient illumination changes for stable tracking even under such circumstances.

$$\mathbf{I}_t \approx \mathbf{I}_{m,t} + \mathbf{I}_{i,t}. \tag{1}$$

We assume that image intensity changes arise from both motion and illumination variations as shown in (1). $\mathbf{I}_t$ is image gradient with respect to time $t$, and both $\mathbf{I}_{m,t}$ and $\mathbf{I}_{i,t}$ are the instantaneous image intensity changes due to motion and illumination variations respectively.

### *A. Motion*

First, we assume static ambient illumination and thus that instantaneous image intensity changes arise from variations in motion only. If then, the following BCCE holds true.

$$\begin{aligned} I(x,y,t) &= I(x+\Delta x, y+\Delta y, t+\Delta t) \\ &\approx I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t. \end{aligned} \tag{2}$$

$$\frac{\partial I}{\partial x} v_x + \frac{\partial I}{\partial y} v_y = -\frac{\partial I_m}{\partial t}, \tag{3}$$

where $v_x = dx/dt$ and $v_y = dy/dt$ are the x- and y-components of the 2D image velocity $\mathbf{v}$ of object motion after projection onto the image plane. In addition, we replace $\partial I / \partial t$ with $\partial I_m / \partial t$ to denote that the intensity changes are due to motion variations.

$$\begin{bmatrix} I_x & I_y \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} = -I_{m,t}, \tag{4}$$

where $I_x$, $I_y$, and $I_{m,t}$ are the spatial and temporal derivatives of the image intensity computed at location $\mathbf{p} = \begin{bmatrix} x & y \end{bmatrix}^T$ respectively, where $I_{m,t}$ arise from the motion changes. Under the perspective projection camera model with focal length $f$, 2D image velocities can be related to 3D object velocities by the following equations.

$$\mathbf{v} = \begin{bmatrix} v_x \\ v_y \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} f & 0 & -x \\ 0 & f & -y \end{bmatrix} \mathbf{V}, \tag{5}$$

where $\mathbf{V} = \begin{bmatrix} V_X & V_Y & V_Z \end{bmatrix}^T$ is the 3D velocity of a point $\mathbf{P} = \begin{bmatrix} X & Y & Z \end{bmatrix}^T$, corresponding to the image pixel $\mathbf{p}$, in the camera coordinate frame.

Any rigid body motion can be expressed in terms of the instantaneous rotations and translation of the object. For small inter-frame rotations, the rotation matrix can be linearly approximated as ( $\Delta\mathbf{R} \approx \mathbf{I} + [\Delta\mathbf{r}]_\times$ ) by the angle-axis formula. $\mathbf{I}$ is a $3\times3$ identity matrix, and $[]_\times$ denotes a skew-symmetric matrix. Also, assuming that time interval $\Delta t$ is unity, temporal derivatives of rotation and translation vectors can be approximated by finite differences $\Delta\mathbf{r}$, $\Delta\mathbf{t}$ respectively.

$$\mathbf{V} \approx \mathbf{R} \begin{bmatrix} \mathbf{I} & -[\mathbf{P}_o]_\times \end{bmatrix} \begin{bmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{r} \end{bmatrix}, \tag{6}$$

where $\mathbf{P}_o$ is a 3D sampled model point in the object coordinate frame corresponding to the point $\mathbf{P}$ in the camera reference frame. $\mathbf{R}$ is the rotation matrix computed in the previous frame between the camera and object coordinate frames. $\Delta\mathbf{r}$ and $\Delta\mathbf{t}$ are the inter-frame rotation and translation vectors expressed in the object coordinate frame, respectively. Substituting (5) and (6) into (4), we obtain a simple linear equation as shown below.

$$\frac{1}{Z} \begin{bmatrix} fI_x & fI_y & -\left(xI_x + yI_y\right) \end{bmatrix} \mathbf{R} \begin{bmatrix} \mathbf{I} & -[\mathbf{P}_o]_\times \end{bmatrix} \begin{bmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{r} \end{bmatrix} = -I_{m,t}. \tag{7}$$

Because (7) is linear with respect to motion parameters, we can combine it across $n$ pixels by stacking the equations in matrix form. $n$ is the number of model points that can be seen from the camera under the current estimated head pose.

$$\begin{pmatrix} \frac{1}{Z_1} \begin{bmatrix} fI_{x,1} & fI_{y,1} & -\left(x_1 I_{x,1} + y_1 I_{y,1}\right) \end{bmatrix} \mathbf{R} \begin{bmatrix} \mathbf{I} & -[\mathbf{P}_{o,1}]_\times \end{bmatrix} \\ \vdots \\ \frac{1}{Z_n} \begin{bmatrix} fI_{x,n} & fI_{y,n} & -\left(x_n I_{x,n} + y_n I_{y,n}\right) \end{bmatrix} \mathbf{R} \begin{bmatrix} \mathbf{I} & -[\mathbf{P}_{o,n}]_\times \end{bmatrix} \end{pmatrix} \begin{pmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{r} \end{pmatrix} = \begin{pmatrix} -I_{m,t,1} \\ \vdots \\ -I_{m,t,n} \end{pmatrix}. \tag{8}$$

Let the left-hand side of (8) be $\mathbf{M}$ and the right-hand side be $\mathbf{I}_{m,t}$. Then, (8) can be represented in compact matrix form as shown below.

$$\mathbf{M}\boldsymbol{\alpha} = \mathbf{I}_{m,t},\ \boldsymbol{\alpha} = \begin{bmatrix} \Delta\mathbf{t} & \Delta\mathbf{r} \end{bmatrix}^T. \tag{9}$$

### *B. Illumination*

As mentioned in the beginning of Section II, BCCE does not hold true under time-varying illumination conditions. To handle face image variations due to changes in lighting conditions, many methods approximate the intensity changes due to illumination variations as a linear combination of illumination bases that are obtained from the training samples of different people taken under a wide variety of lighting conditions [7]. However, these kinds of subspace-based methods construct an illumination subspace from training images for different people, which includes not only illumination conditions but also face identities. This subspace is not capable of representing the lighting conditions uniquely, because the intrinsic (facial geometry and albedo) and the extrinsic (illumination conditions) information is mixed. Otherwise, extremely large training sets would be needed. Furthermore, these methods need a prior training process and thus suffer from the cost of training data acquisition and processing.

Hence, in this paper, by computing these illumination bases online from the registered face images, after estimating the head poses, user-specific illumination bases can be obtained, and therefore illumination-robust tracking without a prior learning process can be possible as shown in Fig. 1. Therefore, we can approximate the intensity changes due to illumination variations as a linear combination of illumination bases obtained through online illumination modeling based on principal component analysis (PCA) as shown below.

$$\mathbf{L}\boldsymbol{\beta} = \mathbf{I}_{i,t}, \tag{10}$$

where $\mathbf{I}_{i,t}$ is the instantaneous image intensity changes due to illumination variations. The columns of the matrix $\mathbf{L} = \left[\mathbf{l}_1, \cdots, \mathbf{l}_k\right]$ are the illumination bases obtained by PCA, and $\boldsymbol{\beta}$ is the illumination coefficient vector. $k$ is the number of principal components.

### C. Combined into Unified Stereo-Based Framework

First, BCCE for each left and right camera of a stereo-rig can be derived in the same way as (8) and (9) in the single camera system.

$$\begin{bmatrix} \frac{1}{Z_{l,1}}\left[f_l I_{x,l,1} \quad f_l I_{y,l,1} \quad -\left(x_{l,1} I_{x,l,1} + y_{l,1} I_{y,l,1}\right)\right]\mathbf{R}_l\left[\mathbf{I} \quad -\left[\mathbf{P}_{o,l,1}\right]_\times\right] \\ \vdots \\ \frac{1}{Z_{l,n_l}}\left[f_l I_{x,l,n_l} \quad f_l I_{y,l,n_l} \quad -\left(x_{l,n_l} I_{x,l,n_l} + y_{l,n_l} I_{y,l,n_l}\right)\right]\mathbf{R}_l\left[\mathbf{I} \quad -\left[\mathbf{P}_{o,l,n_l}\right]_\times\right] \end{bmatrix} = \mathbf{M}_l,$$

$$\begin{bmatrix} \frac{1}{Z_{r,1}}\left[f_r I_{x,r,1} \quad f_r I_{y,r,1} \quad -\left(x_{r,1} I_{x,r,1} + y_{r,1} I_{y,r,1}\right)\right]\mathbf{R}_r\left[\mathbf{I} \quad -\left[\mathbf{P}_{o,r,1}\right]_\times\right] \\ \vdots \\ \frac{1}{Z_{r,n_r}}\left[f_r I_{x,r,n_r} \quad f_r I_{y,r,n_r} \quad -\left(x_{r,n_r} I_{x,r,n_r} + y_{r,n_r} I_{y,r,n_r}\right)\right]\mathbf{R}_r\left[\mathbf{I} \quad -\left[\mathbf{P}_{o,r,n_r}\right]_\times\right] \end{bmatrix} = \mathbf{M}_r, \tag{11}$$

$$\mathbf{I}_{m,t,l} = \left[-I_{m,t,l,1} \ldots - I_{m,t,l,n_l}\right]^T, \mathbf{I}_{m,t,r} = \left[-I_{m,t,r,1} \ldots - I_{m,t,r,n_r}\right]^T, \tag{12}$$

where $n_l$ and $n_r$ are the number of 3D sampled model points that can be seen from the left and right cameras under the current estimated head pose respectively.

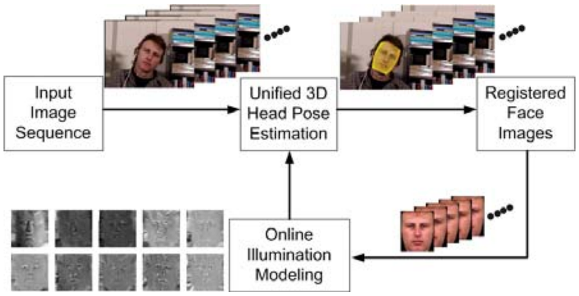

Fig. 1. Online illumination modeling based on PCA.

After combining the above equations into the stereo-based framework, we can obtain a simple linear equation with respect to inter-frame motion parameter $\boldsymbol{\alpha}$ as shown below.

$$\begin{bmatrix} \mathbf{M}_l \\ \mathbf{M}_r \end{bmatrix}\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{I}_{m,t,l} \\ \mathbf{I}_{m,t,r} \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{r} \end{bmatrix}. \tag{13}$$

In the same way as in Section II.B, we can also model the instantaneous intensity changes due to illumination variations as a linear combination of illumination bases for each left and right face image as shown below.

$$\begin{bmatrix} \mathbf{L}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_r \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta}_l \\ \boldsymbol{\beta}_r \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{i,t,l} \\ \mathbf{I}_{i,t,r} \end{bmatrix}, \tag{14}$$

where $\mathbf{L}_l = \left[\mathbf{l}_{l,1}, \ldots, \mathbf{l}_{l,k}\right]$ and $\mathbf{L}_r = \left[\mathbf{l}_{r,1}, \ldots, \mathbf{l}_{r,k}\right]$ are two sets of illumination bases for the left and right face images respectively, which are obtained by removing the rows of $\mathbf{L}$ corresponding to invisible model points from each left and right camera under the current estimated head pose. $\mathbf{L}$ is computed through online illumination modeling based on PCA from both the left and right registered face images that had been stored until the previous frame. $k \le 2F - 1$ is the number of illumination bases, and $F$ is the number of frames. $\boldsymbol{\beta}_l$ and $\boldsymbol{\beta}_r$ are the illumination coefficient vectors for the left and right face images respectively. $\mathbf{I}_{i,t,l}$ and $\mathbf{I}_{i,t,r}$ are the instantaneous image intensity changes due to illumination variations for the left and right face images respectively.

Because we assumed (1) in the beginning of Section II, and because (13) and (14) are linear with respect to motion parameter $\boldsymbol{\alpha}$ and illumination coefficient vectors $\boldsymbol{\beta}_l$ and $\boldsymbol{\beta}_r$ respectively, we can combine them into a unified stereo-based framework as shown below.

$$\begin{bmatrix} \mathbf{M}_l & \mathbf{L}_l & \mathbf{0} \\ \mathbf{M}_r & \mathbf{0} & \mathbf{L}_r \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}_l \\ \boldsymbol{\beta}_r \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{t,l} \\ \mathbf{I}_{t,r} \end{bmatrix}. \tag{15}$$

Let the left-hand side of (15) be $\mathbf{A}$ and the right-hand side be $\mathbf{b}$. Then, the weighted least-squares solution of (15) can be easily obtained as shown below.

$$\begin{aligned} \mathbf{s}^* &= \arg\min_{\mathbf{s}} \left\|\mathbf{WAs} - \mathbf{Wb}\right\|^2 \\ &= \left(\left(\mathbf{WA}\right)^T\left(\mathbf{WA}\right)\right)^{-1}\left(\mathbf{WA}\right)^T\left(\mathbf{Wb}\right), \end{aligned} \tag{16}$$

where $\mathbf{W}$ is a diagonal matrix whose components are pixel weights assigned according to their projection densities as in [3]. Finally, motion parameters between the object and camera coordinate frames are updated by (17) and iterated until the estimates of the parameters converge. Initial motion parameters are assumed to be known.

$$\begin{aligned} &\mathbf{R}_l \leftarrow \mathbf{R}_l \Delta \mathbf{R}, \;\; \mathbf{T}_l \leftarrow \mathbf{R}_l \Delta \mathbf{t} + \mathbf{T}_l, \\ &\mathbf{R}_r \leftarrow \mathbf{R}_r \Delta \mathbf{R}, \;\; \mathbf{T}_r \leftarrow \mathbf{R}_r \Delta \mathbf{t} + \mathbf{T}_r. \end{aligned} \qquad (17)$$

where $\mathbf{R}_r$ and $\mathbf{T}_r$ are related to $\mathbf{R}_l$ and $\mathbf{T}_l$ through the stereo geometry as $\mathbf{R}_r = \mathbf{R}_s^T \mathbf{R}_l$ and $\mathbf{T}_r = \mathbf{R}_s^T (\mathbf{T}_l - \mathbf{T}_s)$.

## III. Face Recognition

As mentioned in Section I, if we can align the face images with pose variations into stabilized views, the recognition task would be much easier, and higher recognition rate can be achieved. Fig. 2 presents how to obtain a pose-compensated face image when given 3D pose information of the head under perspective projection. The general idea of stabilization is as follows. First, we can estimate the current pose of the ellipsoid corresponding to a human head using the proposed unified motion estimation technique. If so, then we can find out the relationship between all surface points of the ellipsoid and their projections onto the input image plane under the perspective projection model. Second, by projecting all surface points onto the $X_o Y_o$ plane, we can generate a stabilized view image. Following this procedure, we can find out the complete relationship between an arbitrary input face image with pose variation and its corresponding stabilized texture map.

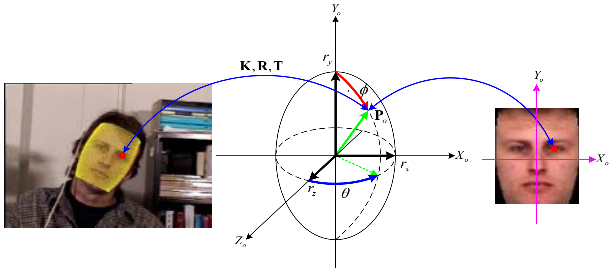

Fig. 2. Geometrical mapping from an input face image to its stabilized texture map under the estimated head pose and perspective projection. **K** represents the camera intrinsic parameter and is assumed to be known. **R** and **T** are the estimated rotation matrix and the translation vector of the current head pose, respectively.

However, there might be missing pixels in the stabilized texture map, which correspond to invisible regions from the camera due to self-occlusion and camera's viewing direction. Therefore, such invisible regions are considered as missing pixels, and their intensities are set to be zeros in the stabilized texture map. This may deteriorate the recognition performance. Therefore, we can also generate a mirrored texture map through a simple mirror operation on the stabilized texture map around its vertical axis as shown in Fig. 3. By doing so, we can make up for the missing pixels and improve the recognition performance.

Finally, simple and efficient frontal face recognition can be easily carried out in the stabilized (or mirrored) texture map space, which is nearly linear-separable, instead of the original input image space that is highly nonlinear and complex.

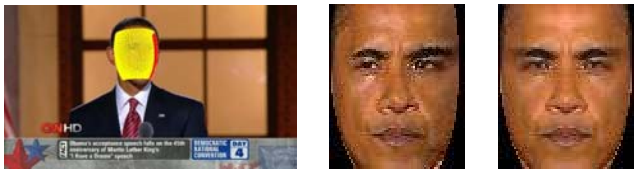

Fig. 3. Mirrored texture map generation. An input image and its stabilized and mirrored texture maps are shown from left to right, respectively.

## IV. Experimental Results

To verify the feasibility and applicability of our proposed 3D head-tracking framework, we performed extensive experiments with three sets of challenging image sequences. All the three experiment sets of stereo image sequences were collected with a stereo vision module named "Bumblebee". All the image sequences were digitized at 30 frames per second at a resolution of $320 \times 240$. Ground truth data for the first and second sets was simultaneously collected via a 3D magnetic sensor named "Flock of Birds". The magnetic sensor has a positional accuracy of 2.54*mm* and rotational accuracy of $0.5^\circ$. The first set consists of 20 image sequences (two sequences for each of 10 subjects) taken under near-uniform illumination conditions. The second set consists of 20 image sequences (two sequences for each of 10 subjects) taken under time-varying illumination. All the sequences in the first and second sets are 300 frames long and are including free and large head motions. The third set was collected for face recognition test and consists of 17 image sequences (16 males + 1 female) taken under near-uniform lighting conditions. All the sequences in this set are 200 frames long and are also including free and large head motions.

Note that all the measured ground truth and the estimates of the visual tracking are expressed with respect to the initial object (head) coordinate frame for the comparison of estimation errors.

### *A. Experiment 1: Near-Uniform Illumination*

The first experiment was designed to compare the performance of the proposed tracker with that of a conventional head tracking with a single camera and also intended to evaluate the effects of online illumination correction. 20 stereo image sequences taken under near-uniform illumination were used in this experiment. Left images of a stereo camera were used for the single camera-based tracker. In this experiment, for modeling the illumination changes in face images, we used 10 illumination bases. They were obtained through online illumination modeling based on PCA from both the left and right registered face images that had been stored until the previous frame.

Fig. 4 presents typical tracking results on one of the test sequences from the first experiment set. The estimations for 3D motion on this sequence are displayed in Fig. 5. This sequence involves large pitch, yaw, and roll motions up to $40^\circ$, $70^\circ$, and $35^\circ$ respectively. "Single" denotes conventional single camera-based tracking defined by (9). "Stereo" represents stereo-based tracking described by (13). This is a simple extension of "Single" to a stereo framework, but not including illumination correction. "Unified stereo" means our proposed unified stereo-based tracking including online illumination correction.

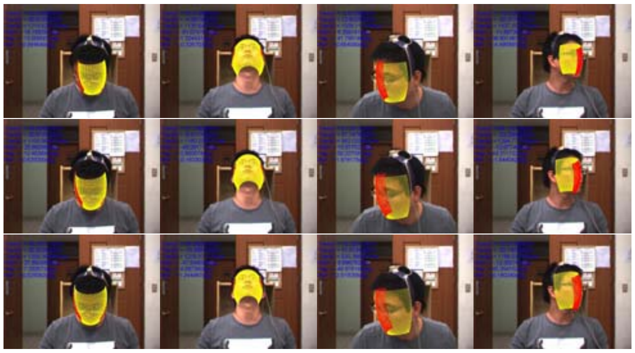

Fig. 4. Typical tracking results on one of the sequences taken under near-uniform illumination. Frames 116, 138, 210, and 251 are shown (left to right). Row 1: single; Row 2: stereo; Row 3: unified stereo.

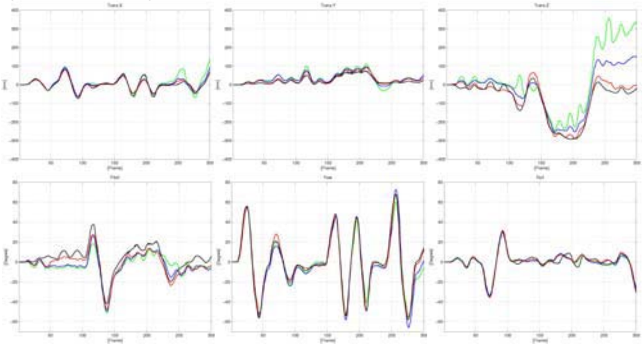

Fig. 5. Comparison between the ground truth and the estimated head poses on the sequence corresponding to Fig. 4. Green line: single; Blue line: stereo; Red line: unified stereo; Black line: the ground truth.

TABLE I
MOTION ESTIMATION ERRORS ON 20 IMAGE SEQUENCES TAKEN UNDER NEAR-UNIFORM ILLUMINATION CONDITIONS

| | Single | Stereo | Ours |
|---|---|---|---|
| Trans. X [mm] | 11.27 | 8.24 | 5.83 |
| Trans. Y [mm] | 9.65 | 6.75 | 4.30 |
| Trans. Z [mm] | 66.61 | 38.62 | 12.19 |
| Pitch [degree] | 5.46 | 3.92 | 2.50 |
| Yaw [degree] | 6.08 | 4.95 | 3.62 |
| Roll [degree] | 2.54 | 2.27 | 1.80 |

Average errors of 3D motion estimation on 20 image sequences are shown in Table I. As can be seen in these results, single camera-based tracking is not robust to large out-of-plane rotations (especially for pitch and yaw) and translation in-depth. A simple extension to stereo-based tracking improves the performance of the tracker to some degree, but there still exist significant tracking errors. On the other hand, even though there are no changes in ambient illumination, motion estimation is greatly improved through the proposed unified stereo-based tracking including online illumination correction compared to stereo-based tracking. This is because self-shading is likely to occur in face images even under uniform illumination, depending on the current head pose. Hence, our proposed unified stereo-based tracking can provide robust motion estimation by reducing the negative effects of self-shading, thanks to the illumination correction term.

### B. Experiment 2: Time-Varying Illumination

The second experiment was set up to evaluate the performance of the proposed tracker under time-varying illumination conditions. In this experiment, we also used 10 illumination bases obtained through online illumination modeling as in Experiment 1.

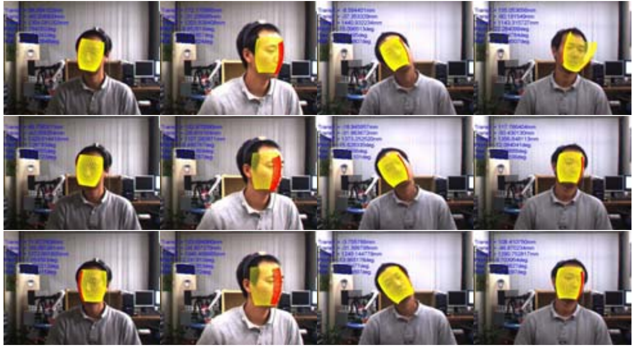

Fig. 6. Typical tracking results on one of the sequences taken under time-varying illumination. Frames 149, 181, 245, and 300 are shown (left to right). Row 1: single; Row 2: stereo; Row 3: unified stereo.

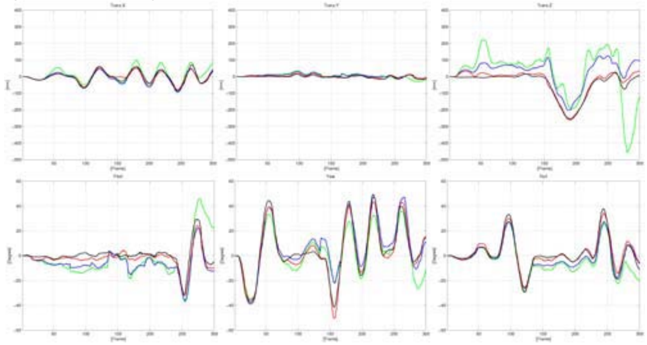

Fig. 7. Comparison between the ground truth and the estimated head poses on the sequence corresponding to Fig. 6. Green line: single; Blue line: stereo; Red line: unified stereo; Black line: the ground truth.

TABLE II
MOTION ESTIMATION ERRORS ON 20 IMAGE SEQUENCES TAKEN UNDER TIME-VARYING ILLUMINATION CONDITIONS

| | Single | Stereo | Ours |
|---|---|---|---|
| Trans. X [mm] | 18.85 | 15.44 | 5.73 |
| Trans. Y [mm] | 16.02 | 10.86 | 4.75 |
| Trans. Z [mm] | 112.37 | 52.68 | 14.91 |
| Pitch [degree] | 9.91 | 7.07 | 3.32 |
| Yaw [degree] | 18.89 | 14.60 | 3.61 |
| Roll [degree] | 6.86 | 6.42 | 2.05 |

Fig. 6 presents typical tracking results on one of the test sequences from the second experiment set. The estimations for 3D head motion on this sequence are displayed in Fig. 7. This sequence also involves large pitch, yaw, and roll motions up to 30°, 50°, and 38° respectively. Whenever there are changes in illumination, significant tracking errors occur in "Single" and "Stereo" tracking. On the other hand, the proposed unified stereo-based tracker shows stable tracking even under time-varying illumination.

Average errors of 3D motion estimation on 20 image sequences are shown in Table II. There exist much larger tracking errors in "Single" and "Stereo" tracking than those in Experiment 1, because they cannot cope with illumination changes. On the other hand, our tracker shows slightly deteriorated but almost similar performance of motion estimation to that evaluated in Experiment 1 even under time-varying illumination, thanks to the illumination correction term.

### C. *Experiment 3: Face Recognition*

The third experiment was intended to verify that our proposed head tracking method is helpful to improve the performance of face recognition. In this experiment, we constructed three test sets such as unregistered, stabilized, and mirrored sets. For the unregistered test set, we manually cropped 200 pairs of stereo face images from the input image sequence for each of 17 classes. For the stabilized set, we obtained 200 pairs of stereo face images registered into frontal views by the proposed unified stereo-based tracker for each class. For the mirrored set, we made a mirror operation on the stabilized test set. For face recognition on each test set, we used only a pair of stereo face images (frontal views) for the training and 200 pairs of stereo face images obtained by the aforementioned methods for the test from each class. For the comparison of performance, we used three linear subspace-based classification methods such as PCA, PCA+LDA (PCA followed by LDA), and DCV [8]-[10].

TABLE III
PERFORMANCE OF FACE RECOGNITION ON OUR LABORATORY TEST SETS USING THREE LINEAR SUBSPACE-BASED CLASSIFICATION METHODS

| | Unregistered | Stabilized | Mirrored |
|---|---|---|---|
| PCA | 62.50% | 84.94% | 92.32% |
| PCA+LDA | 69.35% | 88.06% | 96.12% |
| DCV | 66.24% | 83.29% | 91.15% |

Table III shows the recognition rates with three linear classification methods. As can be seen in the recognition rates, we can verify that face registration is helpful to improve the recognition performance, and also the recognition rate using the mirrored texture maps is much better.

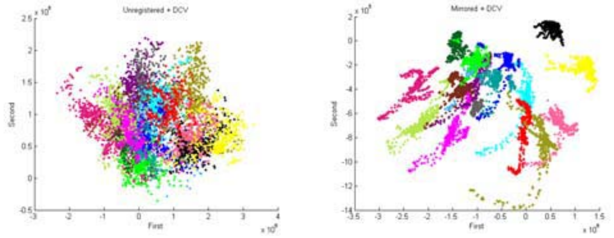

Fig. 8. Distributions of 3400 pairs of test samples from 17 classes, projected onto the two-dimensional subspace spanned by two optimal projection vectors obtained by DCV, for the unregistered and mirrored test sets.

Fig. 8 shows the distributions of 3400 pairs of test samples from 17 classes projected onto the two-dimensional subspace spanned by two optimal projection vectors obtained by DCV for the unregistered and mirrored test sets respectively. As can be seen in this figure, the registered set with mirrored texture maps is well-clustered compared with the unregistered one. Therefore, when using the stabilization scheme based on our proposed 3D head pose estimation, linear classification can be easily applied, and also much higher recognition rate can be achieved than the unregistered case.

## V. CONCLUSION

In this paper, we presented a long-term stable and robust technique for 3D head tracking even in the presence of time-varying illumination conditions. We extended conventional head tracking with a single camera to a stereo-based framework. This partially enables us to cope with large out-of-plane rotations and translation in-depth. In addition, we incorporated the online illumination correction term into this stereo-based framework for more robust motion estimation. We approximated the intensity changes in face images due to illumination variations as a linear combination of illumination bases. Also, by computing these illumination bases online from the registered face images, after estimating the head pose, user-specific illumination bases can be obtained, and finally illumination-robust tracking without a prior learning process that needs a great cost of training data acquisition and processing can be possible.

This paper has shown the feasibility and applicability of the proposed approach by carrying out three challenging experiments. First, it was verified that the proposed unified stereo-based tracking method is able to cope with fast and large out-of-plane rotations and translation in-depth. This is true even under time-varying illumination conditions. Second, it was proved that our proposed unified stereo-based head tracking is helpful to improve the performance of face recognition (over 91% recognition rate when using mirrored texture maps).

## REFERENCES

[1] X. Lu, A. K. Jain, and D. Colby, “Matching 2.5d face scans to 3d models,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 1, 2006.

[2] M. L. Cascia and S. Sclaroff, “Fast, reliable head tracking under varying illumination,” *Int’l Conf. Computer Vision and Pattern Recognition*, 1999.

[3] J. Xiao, T. Kanade, and J. F. Cohn, “Robust full-motion recovery of head by dynamic templates and re-registration techniques,” *Int’l Conf. Automatic Face and Gesture Recognition*, 2002.

[4] V. Blanz and T. Vetter, “Face recognition based on fitting 3d morphable model,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 9, pp. 1063-1074, 2003.

[5] K. H. An and M. J. Chung, “3d head tracking and pose-robust 2d texture map-based face recognition using a simple ellipsoid model,” *Int’l Conf. Intelligent Robots and Systems*, pp. 307-312, 2008.

[6] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone, “FRVT 2002: Evaluation Report,” Mar. 2003.

[7] P. Hallinan, “A low-dimensional representation of human faces for arbitrary lighting conditions,” *Int’l Conf. Computer Vision and Pattern Recognition*, pp. 995-999, 1994.

[8] M. Turk and A. P. Pentland, “Eigenfaces for recognition,” *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[10] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, “Discriminant common vectors for face recognition,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 1, 2005.