Unified Stereo-Based 3D Head Tracking Using Online Illumination Modeling

Kwang Ho An
akh@cheonji.kaist.ac.kr
Myung Jin Chung
mjchung@ee.kaist.ac.kr

Department of Electrical Engineering and Computer Science, KAIST, Republic of Korea

Abstract

This paper investigates the estimation of 3D head poses 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.

1 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, active telecommunication, virtual reality, and visual surveillance. In addition, a face image aligned in terms of the recovered head motion would facilitate facial expression analysis and face recognition.

Thus, many approaches to recover 3D head motion have been proposed [2, 3, 7, 9]. 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*. [3] 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 $\times$ 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*. [9] 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 updated dynamically to diminish the effects of self-occlusion and gradual lighting

BMVC 2009 doi:10.5244/C.23.8

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 models can be significant. This inherent modeling error may degrade the accuracy in motion estimation.

Blanz and Vetter [2] 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 out-of-plane rotations and translation in depth.

With consideration of all of these issues, the coverage of this paper is as follows. As in [1], 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 pose, user-specific illumination bases can be obtained, and therefore illumination-robust tracking without a prior learning process can be possible.

2 3D Head Pose Estimation

Generally, image-based tracking is based on the brightness change constraint equation (BCCE) [6]. 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 more 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 Eq. (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.

2.1 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.

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

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

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 $\mathbf{p} = [x \;\; y]^T$ respectively, where $I_{m,t}$ arises from the motion changes. However, we are interested in solving for 3D velocities of object points, which are related to 3D motion parameter estimation. Under the perspective projection camera model with focal length f, 2D image velocities can be related to 3D object velocities by the following equations.

$$v_x = \frac{d}{dt}\left(f\frac{X}{Z}\right) = \left(\frac{f}{Z}V_X - \frac{x}{Z}V_Z\right), v_y = \frac{d}{dt}\left(f\frac{Y}{Z}\right) = \left(\frac{f}{Z}V_Y - \frac{y}{Z}V_Z\right), \tag{5}$$

where $\mathbf{V} = [V_X \;\; V_Y \;\; V_Z]^T$ is the 3D velocity of a point $\mathbf{P} = [X \;\; Y \;\; Z]^T$, corresponding to the image point $\mathbf{p}$, in the camera coordinate frame. The relationship between the two corresponding velocities can be expressed in compact matrix form as shown below.

$$\mathbf{v} = \begin{bmatrix} v_x \\ v_y \end{bmatrix} = \frac{1}{Z}\begin{bmatrix} f & 0 & -x \\ 0 & f & -y \end{bmatrix}\mathbf{V}. \tag{6}$$

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 3×3 identity matrix, and $[\,]_\times$ denotes a skew-symmetric matrix. Also, assuming that time interval Δt is unity, temporal derivatives of rotation and translation vectors can be approximated by finite differences $\Delta\mathbf{r}$, $\Delta\mathbf{t}$ respectively.

$$\mathbf{V} \approx \mathbf{R}\begin{bmatrix} \mathbf{I} & -[\mathbf{P}_o]_\times \end{bmatrix}\begin{bmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{r} \end{bmatrix}, \tag{7}$$

where $\mathbf{P}_o$ is a 3D sampled model point in the object reference 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 expressed in the object coordinate frame. The above equation describes the relationship between the 3D object velocity in the camera coordinate frame and inter-frame rigid body motion parameters in the object coordinate frame. Substituting Eqs. (6) and (7) into Eq. (4), we obtain a simple linear equation as shown below.

$$\frac{1}{Z}\begin{bmatrix} fI_x & fI_y & -(xI_x + yI_y) \end{bmatrix}\mathbf{R}\begin{bmatrix} \mathbf{I} & -[\mathbf{P}_o]_\times \end{bmatrix}\begin{bmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{r} \end{bmatrix} = -I_{m,t}. \tag{8}$$

The above single linear equation relates the spatial and temporal image intensity derivatives to rigid body motion parameters under the perspective projection model at a single pixel. Because Eq. (8) is linear with respect to motion parameters, we can combine it across n pixels by stacking the equations in a matrix form. n is the number of model points that can be seen from the camera under the current estimated head pose.

$$\begin{bmatrix} \frac{1}{Z_1}\begin{bmatrix} fI_{x,1} & fI_{y,1} & -(x_1 I_{x,1} + y_1 I_{y,1}) \end{bmatrix} \mathbf{R} \begin{bmatrix} \mathbf{I} & -[\mathbf{P}_{o,1}]_\times \end{bmatrix} \\ \vdots \\ \frac{1}{Z_n}\begin{bmatrix} fI_{x,n} & fI_{y,n} & -(x_n I_{x,n} + y_n I_{y,n}) \end{bmatrix} \mathbf{R} \begin{bmatrix} \mathbf{I} & -[\mathbf{P}_{o,n}]_\times \end{bmatrix} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{r} \end{bmatrix} = \begin{bmatrix} -I_{m,t,1} \\ \vdots \\ -I_{m,t,n} \end{bmatrix}. \tag{9}$$

Let the left-hand side of Eq. (9) be $\mathbf{M}$ and the right-hand side be $\mathbf{I}_{m,t}$. Then, Eq. (9) can be represented in compact matrix form as shown below.

$$\mathbf{M}\boldsymbol{\alpha} = \mathbf{I}_{m,t}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{r} \end{bmatrix}. \tag{10}$$

2.2 Illumination

As mentioned in the beginning of Section 2, BCCE does not hold true under time-varying illumination conditions. To handle face image variations due to changes in lighting conditions, many methods have been proposed in the field of face recognition thus far. Among them, for modeling illumination variations, subspace-based methods have often been used [4, 5, 11]. These kinds of methods model the face image variations due to illumination changes with a low-dimensional linear subspace. They 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. 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 are mixed. Otherwise, extremely large training sets would be needed. Also, these methods need a prior learning process and thus suffer from the cost of training data acquisition and processing.

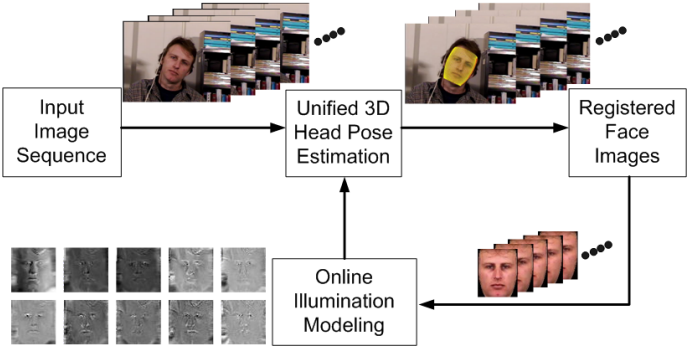

Figure 1: Online illumination modeling.

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) [8] as shown below.

$$\frac{\partial \mathbf{I}_i}{\partial t} = \mathbf{I}_{i,t} = \mathbf{L}\boldsymbol{\beta}, \tag{11}$$

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

2.3 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 Eqs. (9) and (10) in the single camera system.

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

$$\begin{bmatrix} \frac{1}{Z_{r,1}} \begin{bmatrix} f_r I_{x,r,1} & f_r I_{y,r,1} & -\left(x_{r,1} I_{x,r,1} + y_{r,1} I_{y,r,1}\right) \end{bmatrix} \mathbf{R}_r \begin{bmatrix} \mathbf{I} & -[\mathbf{P}_{o,r,1}]_\times \end{bmatrix} \\ \vdots \\ \frac{1}{Z_{r,n_r}} \begin{bmatrix} f_r I_{x,r,n_r} & f_r I_{y,r,n_r} & -\left(x_{r,n_r} I_{x,r,n_r} + y_{r,n_r} I_{y,r,n_r}\right) \end{bmatrix} \mathbf{R}_r \begin{bmatrix} \mathbf{I} & -[\mathbf{P}_{o,r,n_r}]_\times \end{bmatrix} \end{bmatrix} = \mathbf{M}_r, \tag{12}$$

$$\mathbf{I}_{m,t,l} = \begin{bmatrix} -I_{m,t,l,1} \\ \vdots \\ -I_{m,t,l,n_l} \end{bmatrix}, \quad \mathbf{I}_{m,t,r} = \begin{bmatrix} -I_{m,t,r,1} \\ \vdots \\ -I_{m,t,r,n_r} \end{bmatrix}, \tag{13}$$

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.

$$\mathbf{M}_l \boldsymbol{\alpha} = \mathbf{I}_{m,t,l}, \quad \mathbf{M}_r \boldsymbol{\alpha} = \mathbf{I}_{m,t,r}. \tag{14}$$

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

$$\begin{bmatrix} \mathbf{M}_l \\ \mathbf{M}_r \end{bmatrix} \boldsymbol{\alpha} = \begin{bmatrix} \mathbf{I}_{m,t,l} \\ \mathbf{I}_{m,t,r} \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \Delta \mathbf{t} \\ \Delta \mathbf{r} \end{bmatrix}. \tag{15}$$

In the same way as in Section 2.2, 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.

$$\mathbf{L}_l \boldsymbol{\beta}_l = \mathbf{I}_{i,t,l}, \quad \mathbf{L}_r \boldsymbol{\beta}_r = \mathbf{I}_{i,t,r}, \tag{16}$$

where $\mathbf{L}_l = \left[\mathbf{l}_{l,1}, \cdots, \mathbf{l}_{l,k}\right]$ and $\mathbf{L}_r = \left[\mathbf{l}_{r,1}, \cdots, \mathbf{l}_{r,k}\right]$ 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. $k \leq 2F-1$ 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 Eq. (1) in the beginning of Section 2, and because Eqs. (15) and (16) are linear with respect to motion parameters $\boldsymbol{\alpha}$ 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.

$$\begin{bmatrix} \mathbf{M}_l & \mathbf{L}_l & \mathbf{0} \\ \mathbf{M}_r & \mathbf{0} & \mathbf{L}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}_l \\ \boldsymbol{\beta}_r \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{t,l} \\ \mathbf{I}_{t,r} \end{bmatrix}. \tag{17}$$

Let the left-hand side of Eq. (17) be $\mathbf{A}$ and the right-hand side be $\mathbf{b}$. Then, the least-squares solution of Eq. (17) can be easily obtained as shown below.

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \|\mathbf{A}\mathbf{s} - \mathbf{b}\|^2 = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}. \tag{18}$$

Due to the presence of noise, non-rigid motion, occlusion, and projection density, some pixels in the face image may contribute less to motion estimation than others may. To account for these errors, the pixels should be weighted by their contributions. If then, a weighted least-squares solution can be obtained as shown below.

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \|\mathbf{W}\mathbf{A}\mathbf{s} - \mathbf{W}\mathbf{b}\|^2 = \left((\mathbf{W}\mathbf{A})^T(\mathbf{W}\mathbf{A})\right)^{-1}(\mathbf{W}\mathbf{A})^T(\mathbf{W}\mathbf{b}), \tag{19}$$

where $\mathbf{W}$ is a diagonal matrix whose components are pixel weights assigned according to their contributions. Finally, motion parameters between the object and two camera coordinate frames are updated by Eq. (20) and iterated until the estimates of the parameters converge for both the left and right cameras.

$$\mathbf{R}_l \leftarrow \mathbf{R}_l\Delta\mathbf{R}, \quad \mathbf{T}_l \leftarrow \mathbf{R}_l\Delta\mathbf{t} + \mathbf{T}_l, \quad \mathbf{R}_r \leftarrow \mathbf{R}_r\Delta\mathbf{R}, \quad \mathbf{T}_r \leftarrow \mathbf{R}_r\Delta\mathbf{t} + \mathbf{T}_r, \tag{20}$$

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 $\mathbf{T}_r = \mathbf{R}_s^T(\mathbf{T}_l - \mathbf{T}_s)$.

3 Experimental Results

To verify the feasibility and applicability of our proposed 3D head-tracking framework, we performed extensive experiments with two sets of challenging image sequences. Two experiment sets of stereo image sequences were collected with a stereo vision module named "Bumblebee". Ground truth data was simultaneously collected via a 3D magnetic sensor named "Flock of Birds". All the stereo image sequences were digitized at 30 frames per second at a resolution of 320×240. The magnetic sensor has a positional accuracy of 2.54 *mm* and rotational accuracy of 0.5°. The first set consists of 20 stereo image sequences (two

sequences for each of 10 subjects) taken under near-uniform illumination conditions. The second set consists of 20 stereo image sequences (two sequences for each of 10 subjects) taken under time-varying illumination. All the sequences in the two experiment sets are 300 frames long and are 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 coordinate frame for the comparison of estimation errors.

3.1 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. 2 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. 3. This sequence involves large pitch, yaw, and roll motions up to 40°, 70°, and 35° respectively.

"Single" denotes conventional single camera-based tracking defined by Eq. (10). "Stereo" represents stereo-based tracking described by Eq. (15). 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.

Average errors of 3D motion estimation on 20 image sequences are shown in Table 1. 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 direction. 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 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.

3.2 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. 4 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. 5. 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 2. As can be seen in Table 2, 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

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

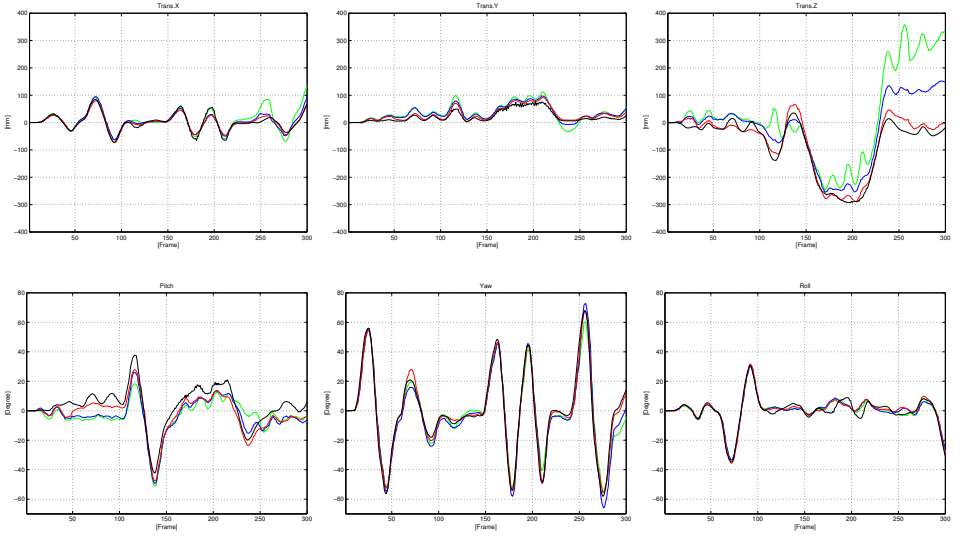

Figure 3: Comparison between the ground truth and the estimated head poses on the sequence corresponding to Fig. 2. Green line: single; Blue line: stereo; Red line: unified stereo; Black line: the ground truth.

	T(x)[mm]	T(y)[mm]	T(z)[mm]	R(x)[°]	R(y)[°]	R(z)[°]
Sinlge	11.27	9.65	66.61	5.46	6.08	2.54
Stereo	8.24	6.75	38.62	3.92	4.95	2.27
Unified stereo	5.83	4.30	12.19	2.50	3.62	1.80

Table 1: Motion estimation errors on 20 image sequences taken under near-uniform illumination conditions.

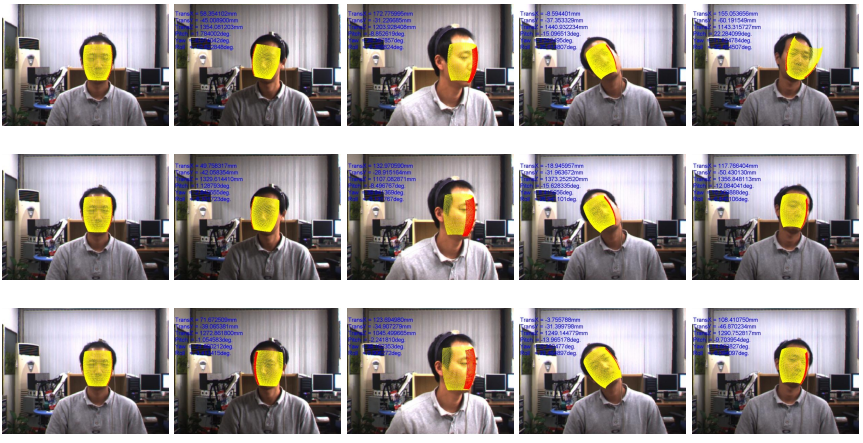

Figure 4: Typical tracking results on one of the sequences taken under time-varying illumination. Frames 1, 149, 181, 245, and 300 are shown from left to right. Row 1: single; Row 2: stereo; Row 3: unified stereo.

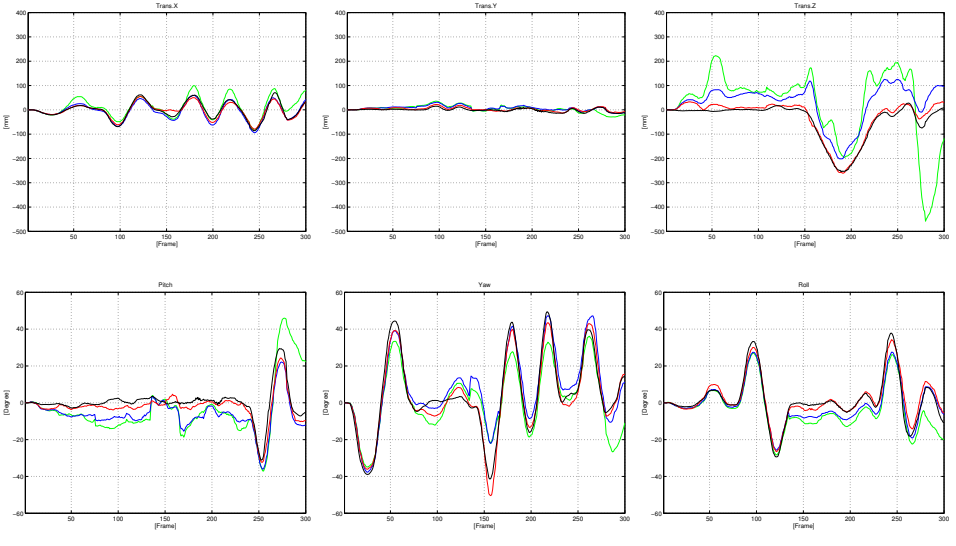

Figure 5: Comparison between the ground truth and the estimated head poses on the sequence corresponding to Fig. 4. Green line: single; Blue line: stereo; Red line: unified stereo; Black line: the ground truth.

	T(x)[mm]	T(y)[mm]	T(z)[mm]	R(x)[°]	R(y)[°]	R(z)[°]
Sinlge	18.85	16.02	112.37	9.91	18.89	6.86
Stereo	15.44	10.86	52.68	7.07	14.60	6.42
Unified stereo	5.73	4.75	14.91	3.32	3.61	2.05

Table 2: Motion estimation errors on 20 image sequences taken under time-varying illumination conditions.

other hand, our unified stereo-based tracker shows almost similar performance of motion estimation to that evaluated in Experiment 1 even under time-varying illumination, thanks to the online illumination correction term.

4 Conclusion

In this paper, we presented a long-term stable and robust technique for 3D head tracking even in the presence of 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 illumination correction 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 can be possible. Extensive experiments using the ground truth have shown 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.

References

[1] Kwang Ho An and Myung Jin Chung. 3d head tracking and pose-robust 2d texture map-based face recognition using a simple ellipsoid model. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 307–312, Sept. 2008.

[2] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, Sept. 2003.

[3] Marco La Cascia, Stan Sclaroff, and Vassilis Athitsos. Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3d models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):322–336, Apr. 2000.

[4] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, Jun 2001.

[5] Peter W. Hallinan. A low-dimensional representation of human faces for arbitrary lighting conditions. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pages 995–999, Jun 1994.

[6] Berthold K. P. Horn and E. J. Weldon Jr. Direct methods for recovering motion. *International Journal of Computer Vision*, 2(1):51–76, 1988.

[7] Shay Ohayon and Ehud Rivlin. Robust 3d head tracking using camera pose estimation. In *Proc. IEEE International Conference on Pattern Recognition*, pages 1063–1066, 2006.

[8] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[9] Jing Xiao, Takeo Kanade, and Jeffrey F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 156–162, May 2002.

[10] Alan L. Yuille, Daniel Snow, Russell A. Epstein, and Peter N. Belhumeur. Determining generative models of objects under varying illumination: Shape and albedo from multiple images using svd and integrability. *International Journal of Computer Vision*, 35(3):203–222, 1999.