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

An accurate estimation of 3D head position and orientation is important in many applications. Thus, many approaches to recover 3D head motion have been proposed [2, 4, 6]. However, 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. Also, most of the existing approaches are not considering illumination correction explicitly.

With consideration of all of these issues, the coverage of this paper is as follows. 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. Also, 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.

Generally, image-based tracking is based on the brightness change constraint equation (BCCE) [3]. 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 assume that image intensity changes arise from both motion and illumination variations as shown in Eq. (1).

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

where $\mathbf{I}_t$ 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.

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 as in [1].

$$\mathbf{M}\boldsymbol{\alpha} = \mathbf{I}_{m,t}, \quad \boldsymbol{\alpha} = [\Delta\mathbf{t} \;\; \Delta\mathbf{r}]^T, \tag{2}$$

$$\mathbf{M} = \begin{pmatrix} \frac{1}{Z_1}\begin{bmatrix} fI_{x,1} & fI_{y,1} & -(x_1 I_{x,1} + y_1 I_{y,1}) \end{bmatrix}\mathbf{R}\begin{bmatrix} \mathbf{I} & -[\mathbf{P}_{o,1}]_\times \end{bmatrix} \\ \vdots \\ \frac{1}{Z_n}\begin{bmatrix} fI_{x,n} & fI_{y,n} & -(x_n I_{x,n} + y_n I_{y,n}) \end{bmatrix}\mathbf{R}\begin{bmatrix} \mathbf{I} & -[\mathbf{P}_{o,n}]_\times \end{bmatrix} \end{pmatrix}, \tag{3}$$

$$\mathbf{I}_{m,t} = \begin{pmatrix} -I_{m,t,1} & \dots & -I_{m,t,n} \end{pmatrix}^T, \tag{4}$$

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. $\mathbf{I}$ is a 3×3 identity matrix, and $[\,]_\times$ denotes a skew-symmetric matrix. $\mathbf{P}_o$ is a 3D sampled model point in the object reference frame corresponding to the point $\mathbf{p}$, and n is the number of model points that can be seen from the camera under the current estimated head pose. $\mathbf{R}$ is the rotation matrix computed in the previous frame between the camera and object coordinate frames. The above linear equation relates the spatial and temporal image intensity derivatives to inter-frame rigid body motion parameters $(\Delta\mathbf{t}, \Delta\mathbf{r})$ under the perspective projection model with focal length f.

As mentioned above, BCCE does not hold true under time-varying illumination conditions. To handle face image variations due to changes in lighting conditions, we model them with a low-dimensional illumination subspace obtained through PCA [5]. Also, 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.

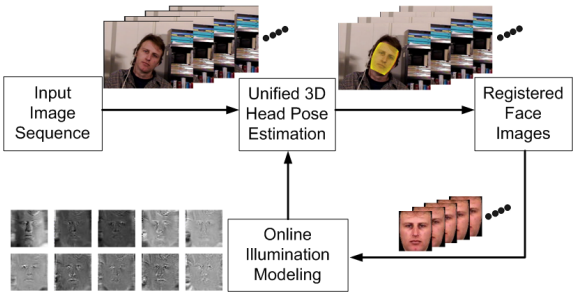

Figure 1: Online illumination modeling.

$$\mathbf{I}_{i,t} = \mathbf{L}\boldsymbol{\beta}, \tag{5}$$

where 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. $\mathbf{I}_{i,t}$ is the instantaneous image intensity changes due to illumination variations.

Finally, we can simply extend Eqs. (2) and (5) to stereo-based frameworks.

$$\mathbf{M}_l\boldsymbol{\alpha} = \mathbf{I}_{m,t,l}, \quad \mathbf{M}_r\boldsymbol{\alpha} = \mathbf{I}_{m,t,r}. \tag{6}$$

$$\mathbf{L}_l\boldsymbol{\beta}_l = \mathbf{I}_{i,t,l}, \quad \mathbf{L}_r\boldsymbol{\beta}_r = \mathbf{I}_{i,t,r}. \tag{7}$$

Because we assumed Eq. (1) in the beginning, and because Eqs. (6) and (7) 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.

$$\begin{bmatrix} \mathbf{M}_l & \mathbf{L}_l & \mathbf{0} \\ \mathbf{M}_r & \mathbf{0} & \mathbf{L}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}_l \\ \boldsymbol{\beta}_r \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{t,l} \\ \mathbf{I}_{t,r} \end{bmatrix}. \tag{8}$$

Let the left-hand side of Eq. (8) be $\mathbf{A}$ and the right-hand side be $\mathbf{b}$. 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. If then, a weighted least-squares solution of Eq. (8) can be obtained as shown below.

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \|\mathbf{WAs} - \mathbf{Wb}\|^2 = \left((\mathbf{WA})^T(\mathbf{WA})\right)^{-1}(\mathbf{WA})^T(\mathbf{Wb}), \tag{9}$$

where $\mathbf{W}$ is a diagonal matrix whose components are pixel weights assigned according to their contributions.

[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. IROS*, pages 307–312, Sept. 2008.

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

[3] Berthold K. P. Horn and E. J. Weldon Jr. Direct methods for recovering motion. *IJCV*, 2(1):51–76, 1988.

[4] 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 T-PAMI*, 22(4):322–336, Apr. 2000.

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

[6] Jing Xiao, Takeo Kanade, and Jeffrey F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. In *Proc. AFGR*, pages 156–162, May 2002.