3D HEAD TRACKING USING NORMAL FLOW CONSTRAINTS IN A VEHICLE ENVIRONMENT

Batu Akan, Müjdat Çetin, Aytül Erçil

Sabanci University
Faculty of Engineering and Natural Sciences
Orhanli - Tuzla, 34956 Istanbul, TURKEY

ABSTRACT

Head tracking is a key component in applications such as human computer interaction, person monitoring, driver monitoring, video conferencing, and object-based compression. The motion of a driver's head can tell us a lot about his/her mental state; e.g. whether he/she is drowsy, alert, aggressive, comfortable, tense, distracted, etc. This paper reviews an optical flow based method to track the head pose, both orientation and position, of a person and presents results from real world data recorded in a car environment.

1. INTRODUCTION

Driver behavior modeling and fatigue detection is an important feature in developing new driver assistance systems and smart cars. These intelligent vehicles are intended to be able to warn or activate other safety measures when hazardous situations have been detected such as fatigued or drunk driver, so that a system can be developed to actively control the driver before he/she becomes too drowsy, tired or distracted [1]. The pose of the head can reveal numerous clues about alertness, drowsiness or whether the driver is comfortable or not. Furthermore knowing the pose of the head will provide a basis for robust facial feature extraction and feature point tracking.

This paper reviews a method for tracking the driver's head using normal flow constraint (NFC) [2] which is an extension of the original optical flow algorithm [3]. Optical flow is the two-dimensional vector field which is the projection of the three-dimensional motion onto an image plane [4]. It is often required to use complex 3D models or non-linear estimation techniques to recover the 3D motion when depth information is not available. However when such observations are available from devices such as laser range finders or stereo cameras, 3D rigid body motion can be estimated using linear estimation techniques. Furthermore combining brightness and depth constraints tend to provide more accuracy for sub pixel movements[2] [5].

In the following section preprocessing and initialization of the tracker, then the derivation of brightness and depth constraints and finally a solution for obtaining motion parameters is described. In section 3,we present experimental results demonstrating that the algorithm can reliably track the drivers head movements using stereo data collected from a dynamic car environment.

2. MOTION ESTIMATION

2.1. Face Detection

At the beginning of motion estimation algorithm, a fast face detector [6] scans the intensity image for face regions. The face detector is trained to detect only frontal faces therefore it can be assumed that the initial rotation of the head is aligned with the camera. The initial region for the head is detected by the face detector and then at each step the region is updated based on the tracker output.

2.2. Brightness Constancy Constraint

A 3D point in space is represented by its coordinate vector $\vec{X} = [X\ Y\ Z]^T$ and the 3D velocity of this point is represented as $\vec{V} = [V_x\ V_y\ V_z]^T$. When this point is projected onto the camera image plane using some projection model, the point will be mapped to 2D image coordinates $\vec{x} = [x\ y]^T$ and the motion of the 3D point in space will induce a corresponding 2D velocity vector onto the camera image plane $\vec{v} = [v_x\ v_y]^T$.

An equation can be derived that relates the change in image brightness at a point to the motion of the brightness pattern. Assume that $I(x, y, t)$ represents the brightness of a point (x, y) in the image plane at a time t. It can be assumed that the brightness of a particular point in the pattern remains constant even though the point has moved. This assumption is only partially true in practice. In situations such as occlusions, disocclusion, changes in intensity due to changes in lighting, the appearance of pixel patches does not represent physical movement of points in space. The assumption may be expressed for frames at t and $t + 1$ as follows:

$$I(x,y,t) = I(x+v_x(x,y,t), y+v_y(x,y,t), t+1) \quad (1)$$

where $I(x,y,t)$ represents the image intensity and $v_x(x,y,t)$ and $v_x(x,y,t)$ are the x and y components of the 2D velocity vector. Taylor expansion of the right hand side of Equation (1) is

$$\begin{aligned} I(x,y,t) = I(x,y,t) + I_x(x,y,t)v_x(x,y,t) \\ +I_y(x,y,t)v_y(x,y,t) + I_t(x,y,t) \end{aligned} \quad (2)$$

where $I_x(x,y,t)$,$I_y(x,y,t)$ and $I_t(x,y,t)$ are the image gradients with respect to x, y and t as a function of space and time. Canceling out the $I(x,y,t)$ terms and rearranging Equation (2) into a matrix form yields the commonly used optical flow equation:

$$-I_t = [I_x \; I_y] \begin{bmatrix} v_x \\ v_y \end{bmatrix} \quad (3)$$

The above equation constraints the velocities in the 2D image plane, but we are interested in the 3D-world velocities. Therefore for a perspective projection camera with focal length f we obtain:

$$\begin{aligned} v_x &= \frac{dx}{dt} = \frac{f}{Z}V_x - \frac{x}{Z}V_z \\ v_y &= \frac{dy}{dt} = \frac{f}{Z}V_y - \frac{y}{Z}V_z \end{aligned} \quad (4)$$

when written in matrix form, becomes:

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} f & 0 & -x \\ 0 & f & -y \end{bmatrix} \begin{bmatrix} V_x \\ V_y \\ V_z \end{bmatrix} \quad (5)$$

By substituting the righthand side of equation (5) for $\vec{v}$ into equation (3), we obtain the brightness constraint equation for 3D object velocities:

$$\begin{aligned} -I_t &= \frac{1}{Z}[I_x \; I_y] \begin{bmatrix} f & 0 & -x \\ 0 & f & -y \end{bmatrix} \vec{V} \\ &= \frac{1}{Z}[fI_x \; fI_y \; -(xI_x + yI_y)]\vec{V} \end{aligned} \quad (6)$$

Any rigid body motion can be expressed as an object going under instantaneous translation $\vec{T} = [t_x \; t_y \; t_z]^T$ and instantaneous rotation $\Omega = [w_x \; w_y \; w_z]^T$ where Ω represents the orientation of the axis of rotation and $|\Omega|$ is the magnitude of rotation per unit time. For small rotations $\vec{V}$ can be approximated as:

$$\vec{V} \approx \vec{T} + \Omega \times \vec{X} = \vec{T} - \vec{X} \times \Omega \quad (7)$$

The cross product of two vectors can be written as the product of a skew-symmetric matrix and a vector. By rearranging $\vec{X} \times \Omega$ into:

$$\vec{X} \times \Omega = \hat{X}\Omega, \text{ where } \hat{X} = \begin{bmatrix} 0 & -Z & Y \\ Z & 0 & -X \\ -Y & X & 0 \end{bmatrix}$$

we can express equation (7) in a matrix form

$$\vec{V} = Q\vec{\phi} \quad (8)$$

where $\vec{\phi} = [\vec{T}^T \; \vec{\Omega}^T]^T$ is the instantaneous motion vector and where

$$Q = [I \; -\hat{X}] = \begin{bmatrix} 1 & 0 & 0 & 0 & -Z & Y \\ 0 & 1 & 0 & Z & 0 & -X \\ 0 & 0 & 1 & -Y & X & 0 \end{bmatrix}$$

When the righthand side of equation (8) is substituted into equation (6), a linear equation which relates pixel intensity values to rigid body motion parameters is obtained for a single pixel.

$$-I_t = \frac{1}{Z}[fI_x \; fI_y \; -(xI_x + yI_y)]\,\mathbf{Q}\vec{\phi} \quad (9)$$

This is the generic brightness constraints used in many of the previous approaches [2] regarding 3D motion and pose tracking. When 3D world coordinates are not known one needs non-linear estimation techniques to solve for the motion. The estimation problem can be simplified to a linear system using 3D models when shape prior of the object being tracked is known. If 3D-world coordinates are available it is relatively easy to solve for this equation system, and non-linearities can be avoided. Not using 3D shape models reduces any errors introduced in the latter class of approaches.

2.3. Depth Constancy Constraint

Assuming that video rate depth information is available for every pixel in the intensity image, similar formulations can be derived using the disparity image. Therefore any changes in the depth image over time can be related to rigid body motion. A point on the rigid body surface, located at (x,y) at time t will be at location $(x+v_x, y+v_y)$ at time $t+1$. The depth values of any particular point at image space and time should remain the same unless the particular point goes under any depth translation between frames t and $t+1$. In mathematical terms, this can be expressed in a way similar to equation (1)

$$Z(x,y,t)+V_z(x,y,t) = V(x+v_x(x,y,t), y+v_y(x,y,t), t+1) \quad (10)$$

Following the same steps that are used to derive the brightness constancy constraint equation, an analogous depth constancy constraint equation can be derived [2]. Rewriting the first order Taylor series expansion of the right hand side of Equation (10) in matrix form we obtain

$$-Z_t = [Z_x \; Z_y] \begin{bmatrix} v_x \\ v_y \end{bmatrix} - V_z \quad (11)$$

By substituting the 3D world velocities into Equation (11) using the perspective projection model yields:

$$-Z_t = [Z_x\ Z_y] \begin{bmatrix} f & 0 & -x \\ 0 & f & -y \end{bmatrix} \vec{V} - V_z \tag{12}$$

Since any rigid body motion can be expressed as an object going under instantaneous translation $\vec{T}$ and instantaneous rotation Ω. Substituting the 3D velocity vector $\vec{V}$ with $Q\vec{\phi}$ as shown previously, produces

$$-Z_t = \frac{1}{Z}[fZ_x\ fZ_y\ -(Z + xZ_x + yZ_y)]\mathbf{Q}\vec{\phi} \tag{13}$$

The derived formulation above is the analogous form of the equation (9) that relates the change in image brightness at a point to the rigid body motion. Brightness constancy constraint equation depends on the assumption that the brightness of a particular point in the pattern remains constant. Since this assumption is only true under some conditions, the outcome is an approximation at best. In contrast Equation(13) makes use of the change in the disparity image, which reflects the true dynamics of the motion. Since the disparity image is not affected by changes in illumination, the depth constancy constraint in Equation (13) yields more accurate results then in Equation (9) in scenarios that involve illumination changes.

2.4. Orthographic Projection

In most cases of interest in this work, camera projection can be modeled as orthographic projection without introducing much error into the system. Such an approach would simplify the constraint equations therefore would reduce the computational load.

Deriving the analogous versions of Equations (9) and (13) is straightforward. All occurrences of image plane x and y are replaced with their real world counterparts X and Y, therefore any real world velocities will be equivalent to their image plane counterparts; $v_x = V_x$ and $v_y = V_y$. Substituting the simplified versions of real world and image plane velocity relations into Equation (3) a much more simpler form of this relation is obtained

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} V_x \\ V_y \\ V_z \end{bmatrix} \tag{14}$$

Inserting the simplified orthographic projection matrix into Equations (5) and (12) yields to the orthogonal projection analogs of the Equations (9) and (13):

$$-I_t = [I_x\ I_y\ 0]\mathbf{Q}\vec{\phi} \tag{15}$$

$$-Z_t = [Z_x\ Z_y\ -1]\mathbf{Q}\vec{\phi} \tag{16}$$

2.5. Shifting the World Coordinate System

Since Euler rotations are defined around the origin, translating 3D coordinates to the centroid $\vec{X_o} = [X_o\ Y_o\ Z_o]$ would increase the numerical stability of the solution. Such a shift in the coordinate system would only affect $\mathbf{Q}$, and the motion parameter vector $\vec{\phi}$ will compensate the for shift. We can rewrite Equation (9) as

$$-I_t = \frac{1}{Z}[fI_x\ fI_y\ -(xI_x + yI_y)]\,\mathbf{Q}'\vec{\phi}' \tag{17}$$

where

$$\mathbf{Q}\prime = \begin{bmatrix} 1 & 0 & 0 & 0 & (Z - Z_o) & -(Y - Y_o) \\ 0 & 1 & 0 & -(Z - Z_o) & 0 & (X - X_o) \\ 0 & 0 & 1 & (Y - Y_o) & -(X - X_o) & 0 \end{bmatrix}$$

and $\vec{\phi} = [\vec{T}\prime^T\ \vec{\Omega}\prime^T]^T$

2.6. Least Squares Solution

In the previous sections brightness and depth constancy constraint formulation are derived. These formulations try to approximate a single pixel's velocity as it undergoes instantaneous translation and rotation. Since these constraint equations are linear they can be stacked up in a matrix formulation $b_I = \mathbf{H_I}\vec{\phi}$ across N pixels which belong to the rigid object that is being tracked. Where $\vec{b_I} \in \Re^{N\times 1}$ is the temporal intensity derivative and $\vec{H_I} \in \Re^{N\times 6}$ is the constraint matrix for brightness values. Vector $\vec{\phi}$ is the motion vector that is to be solved for. A similar formulation can be obtained for depth constraints $b_Z = \mathbf{H_Z}\vec{\phi}$ as well. Given that $N > 6$, the least squares method can be used to solve for the motion parameter vector $\vec{\phi}$ independently for each system. Alternatively we can combine the two equations into a single equation

$$\vec{b} = \mathbf{H}\vec{\phi},\ \text{where } \mathbf{H} = \begin{bmatrix} \mathbf{H_I} \\ \lambda\mathbf{H_D} \end{bmatrix}, \vec{b} = \begin{bmatrix} \vec{b_I} \\ \lambda\vec{b_D} \end{bmatrix} \tag{18}$$

in order to solve for a single vector $\vec{\phi}$, where λ is the scaling factor for depth constraints. In situations where the disparity image is more reliable than the intensity image, such as fast changing illumination conditions, values higher than 1 should be chosen for λ. In other situations where it is known that intensity image is more reliable than the disparity image, values smaller than 1 should be chosen for λ. The least-squares solution for the equation above is:

$$\vec{\phi} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H^T}\vec{b} \tag{19}$$

The least squares solution gives out the motion vector for a set of pixels that belong to the object of interest. These pixels are selected from the images where both intensity and depth images are well defined.

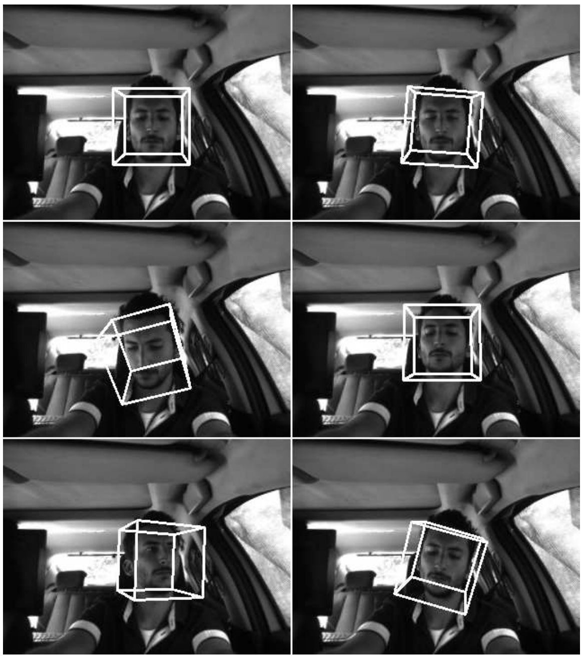

Fig. 1. Result of the driver head tracker at frames: 0, 60, 110, 150, 230, 400

3. PERFORMANCE AND RESULTS

We have tested this algorithm in a real car environment. A bumblebee stereo camera system has been used for data acquisition [7]. The camera hardware analyzes the stereo images and establishes correspondence between pixels in each image. Based on the camera's geometry and the correspondences between pixels in the images, it is possible to determine the distance to points in the scene. Without any special optimizations the tracker can update pose estimations based on 2000-3000 pixels per frame at a rate of 60Hz on a Celeron 1.5 GHz laptop.

Performance of the tracker has been tested using the data collected from "UYANIK" [8]. Several sequences of length 500 frames or roughly 30 seconds of video with both intensity and disparity images has been recorded. The sequences involve all natural head movements: throughout the video the driver rotates his head checking out left, right and rear mirrors of the car and looks down at the gear. Some outputs from the tracking algorithm can be seen in Figure 1.

Due to the differential nature of the tracker over long sequences of video, it is observed that drifts are very likely to occur. These drifts occur because of both noisy image acquisition and that computational errors on these noisy data accumulate over time and result in a drift. Therefore to make the algorithm more robust, it should be backed with an algorithm that resets pose parameters when the drift becomes significant.

4. CONCLUSION AND FUTURE WORK

In this a paper implementation and demonstration of an optical flow based method for tracking a rigid body object, in this work the human head, in a car environment with 6 degrees of freedom is presented. The method is able to handle both translational movements in depth and rotational movements both in and out of the image plane. The algorithm has been successfully tested in a dynamic car environment with sudden changes in illumination.

5. ACKNOWLEDGMENTS

This work was partially supported by the European Commission under Grants FP6-2004-ACC-SSA-2 (SPICE) and MIRG-CT-2006-041919, the Turkish State Planning Organization under the DRIVE-SAFE project, the New Energy and Industrial Technology Development Organization (NEDO) of Japan under the project "International Research Coordination of Driving Behavior Signal Processing based on Large Scale Real World Database"

6. REFERENCES

[1] J.C.Simon, P. Loslever, and P. Popieul, "Using driver's head movements evolution as a drowsiness indicator," in *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, June 2003, pp. 616– 621.

[2] M. Harville, A. Rahimi, T. Darrell, G. G. Gordon, and J. Woodfill, "3d pose tracking with linear depth and brightness constraints," in *International Conference on Computer Vision*, 1999, pp. 206–213.

[3] B. K. P. Horn and V. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–201, 1981.

[4] S. Vedula, S. Baker, R. Collins, T. Kanade, and P. Rander, "Three-dimensional scene flow," in *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, Washington, DC, USA, 1999, p. 722, IEEE Computer Society.

[5] L.P. Morency, A. Rahimi, N. Checka, and T. Darrell, "Fast stereo-based head tracking for interactive environments," in *Proceedings of Conference on Automatic Face and Gesture Recognition*, 2002.

[6] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2002.

[7] Point-Grey Research, "Bumblebee2 stereo vision camera," http://www.ptgrey.com/products/stereo.asp.

[8] H. Abut, H. Erdogan, A. Ercil, B. Cürüklü, H.C. Koman, F. Tas, A.O. Argunsah, S. Cosar, B. Akan, H. Karabalkan, E. Cokelek, R. Ficici, V. Sezer, S. Danis, M. Karaca, M. Abbak. M.G. Uzunbas, K. Ertimen, C. Kalaycioglu, M. Imamoglu, C. Karabat, and M. Peyic, "Data collection with "uyanik": Too much pain; but gains are coming," in *Biennial on DSP for In-Vehicle and Mobile Systems*, June 2007.