3D Pose Tracking with Linear Depth and Brightness Constraints

M. Harville, A. Rahimi,* T. Darrell, G. Gordon, J. Woodfill
Interval Research Corp., 1801C Page Mill Road, Palo Alto, CA 94304
harville,rahimi,trevor,gaile,woodfill@interval.com

This paper appears in *Proceedings of the IEEE International Conference on Computer Vision*, Corfu, Greece. 1999.

Abstract

This paper explores the direct motion estimation problem assuming that video-rate depth information is available, from either stereo cameras or other sensors. We use these depth measurements in the traditional linear brightness constraint equations, and we introduce a new depth constraint equation. As a result, estimation of certain types of motion, such as translation in depth and rotations out of the image plane, becomes more robust. We derive linear brightness and depth change constraint equations that govern the velocity field in 3-D for both perspective and orthographic camera projection models. These constraints are integrated jointly over image regions according to a rigid-body motion model, yielding a single linear system to robustly track 3D object pose. Results are shown for tracking the pose of faces in sequences of synthetic and real images. For a color version of this paper and for video result sequences, see http://www.interval.com/papers/1999-006.

1 Introduction

Estimating 3D object pose is an important problem in computer vision. In particular, it is a challenging aspect of novel human interface applications, which require fast, accurate head or body tracking. Knowledge of users' body position must arrive quickly to adjust the display of an interface in a meaningful, timely manner. For a method to work on body parts with varied clothing or appearance, it should rely on motion of direct image measurements rather than tracking *a priori* features or fixed models.

Previous approaches to pose tracking often rely on assumed models of shape to track motion in 3-D from intensity data. This leads to innaccuracies when the real object deviates from the model, which is typically planar or ellipsoidal.

In this paper we show how we can take advantage of recently developed video-rate range sensors to dramatically improve pose tracking performance. We use models which express parametric motion constraints directly on range and intensity image values, since these methods effectively integrate measurement uncertainty both over image regions and over the motion model. These linearized models yield closed-form solutions, which may be computed quickly.

Our method offers two key innovations to existing direct pose estimation frameworks. First, we use the range information to determine the shape of the object, rather than assume a generic model or estimate structure from motion. This shape is updated with each frame, offering a more accurate representation across time than one provided by an initial or off-line range scan. Second, we derive the depth counterpart to the classic brightness change constraint equation. We use both constraints to jointly solve for motion estimates. Observing the change in depth directly, rather than inferring it from intensity change over time (or subtle perspective effects), can yield more accurate estimates of object motion, particularly for rotation out of the image plane and for translation in depth. Depth information is also less sensitive than intensity data to illumination and shading effects as an object translates and rotates through space, and hence the depth change constraint equation is more reliable than the traditional brightness constraint when these photometric effects are significant.

We use a hardware stereo implementation which offers images of registered depth and intensity at video frame rate. This system relies on the non-parametric census stereo correspondence algorithm [12] and currently runs on a single FPGA PCI card attached to a personal computer [11]. Other real-time range sensing technology may also be used as input to our pose estimation method, as long as registered depth images are available at video rate. "RGBZ" data can directly resolve many of the usual ambiguities present in a single intensity image; in previous papers we have demonstrated the utility of this information for background segmentation [6] and face detection and tracking [5].

The remainder of this paper proceeds as follows. We first summarize previous work on parametric motion methods for pose estimation and head tracking. We then introduce our joint depth and brightness constraint, suitable for image sequences where gradients can be computed both on intensity and depth information. Next, we show how these constraints can be integrated over image regions according

*Currently at the MIT Media Lab.

to a rigid-body motion model. This results in a single linear system with an efficient closed-form solution. We derive the system for both perspective and orthographic camera models. We demonstrate results for tracking objects with known motion in synthetic sequences and for tracking the pose of a user's head in real video sequences.

2 Previous Work

The general problem of estimating object pose from image sequences has been widely studied in the computer vision literature. Here we only outline some of the previous work on this topic, specifically focusing on work in direct parametric motion estimation for head and object tracking.

Rigid and affine models for direct parametric motion estimation have been extensively explored in the past decade. Horn and Weldon provided an early and comprehensive description of the brightness constraints implied by egomotion or the rigid motion of an object in the world [7]. They observed that motion estimation was in general quite difficult to solve with unknown scene depth, although it is possible in several restricted cases. Bergen et al. [1] were among the first to demonstrate image stabilization and object tracking using an affine model with direct image intensity constraints; they utilized a coarse-to-fine algorithm to solve for large motions.

Black and Yacoob [3] applied parametric models to track the motion of a user's head, and also employed non-rigid models to capture expression. For tracking gross head motion they assumed planar face shape, which limits the accuracy and range of motion of their method. Basu and Pentland [2] offered a similar scheme for recovery of rigid motion parameters assuming ellipsoidal shape models and perspective projection. Their method used a precomputed optic flow representation instead of direct brightness constraints. They also represented rigid motion using Euler angles, which can pose certain difficulties at singularities.

More recently, Bregler and Malik [4] introduced the use of the twist representation of rigid motion. Twists, which are commonly used in the field of robotics, are more stable and efficient to compute than Euler angles. They are especially suited to the estimation of chained articulated motion, as Bregler and Malik demonstrated. They estimated twists directly from the image brightness constraint with a scaled orthographic projection model, and they used ellipsoids to model the shape of each limb of the articulated object. To recover motion in depth, they relied on constraints from this articulation and on information from multiple widely-spaced camera views.

Our approach shares similar goals with the derivation of the direct motion stereo equations in Shieh et al.[8], and with the tensor brightness constraint applied to motion stereo shown in Stein and Shashua[9]. However, these methods assume infinitesimal baseline, and rely on a coarse-to-fine solution strategy when used with baselines generating disparities greater than a pixel. Our method uses the range information directly and can be used with any video-rate range sensor, e.g. laser scanner, structured light, or stereo correspondence.

Video-rate range information allows us to express more powerful direct constraints on image and depth gradients, and to linearly estimate pose parameters that can easily track motion in depth. We are able to track the rigid motion of a single unconnected part from a single viewpoint given a monocular sequence of intensity and range imagery.

3 Motion Estimation

We first review the classic brightness change constraint equation and its application to motion estimation under perspective camera projection. We then introduce and develop a second, analogous constraint that operates directly on depth information. Next, we show how to combine these constraints across image pixels into a single linear system which may be solved efficiently for 3-D motion parameters. Finally, we discuss a spatial coordinate shift that greatly improves the motion estimation results.

3.1 The Brightness Change Constraint Equation for Perspective Projection

In all of the following derivations, we will denote the coordinate of a point in 3-D space as $\vec{X} = [\ X \quad Y \quad Z\ ]^T$, and the 3-D velocity of this point as $\vec{V} = [\ V_x \quad V_y \quad V_z\ ]^T$. When we project this point onto the camera image plane via some camera projection model, it will be located at the 2-D image coordinate $\vec{x} = [\ x \quad y\ ]^T$. The 3-D motion of the point in space will induce a corresponding 2-D motion of the projected point in the image plane, and we will express these 2-D velocities as $\vec{v} = [\ v_x \quad v_y\ ]^T$.

The brightness change constraint equation (BCCE) for image velocity estimation arises from the assumption that intensities undergo only local translations from one frame to the next in an image sequence. This assumption is only approximately true in practice, in that it ignores phenomena such as occlusions, disocclusions, and changes in intensity due to changes in lighting. The assumption may be expressed for frames at times t and $t+1$ as follows:

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

$I(x, y, t)$ is the image intensity, and $v_x(x, y, t)$ and $v_y(x, y, t)$ are the x- and y-components of the 2-D velocity field of object motion after projection onto the image plane. If we further assume that the time-varying image intensity is well approximated by a first-order Taylor series expansion, we can expand the right side of the above equation to obtain

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

where $I_x(x,y,t)$, $I_y(x,y,t)$, and $I_t(x,y,t)$ are image intensity gradients with respect to x, y, and t, as a function of space and time. Cancellation of the $I(x,y,t)$ terms and rearrangement into matrix form yields the commonly used gradient formulation of the BCCE [7]:

$$
-I_t = \begin{bmatrix} I_x & I_y \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} \qquad (3)
$$

This equation constrains image plane velocities, but we are interested in solving for 3-D world velocities. For a perspective camera with focal length f, the relationship between the two sets of velocities may be derived from the perspective camera projection equations: $x = \frac{fX}{Z}$, and $y = \frac{fY}{Z}$. Taking the derivatives of these equations with respect to time yields

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

This can be written in matrix form as

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

Substituting the right side of equation (5) for $\vec{v}$ in equation (3), we obtain the BCCE constraint equation in terms of 3-D object velocities:

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

We wish to further constrain the 3-D velocities $\vec{V}$ according to rigid body motion. Any rigid body motion can be expressed in terms of the instantaneous object translation $\vec{T} = [\ t_x \quad t_y \quad t_z\ ]^T$ and the instantaneous rotation of the object about an axis $\vec{\Omega} = [\ \omega_x \quad \omega_y \quad \omega_z\ ]^T$. $\vec{\Omega}$ describes the orientation of the axis of rotation, and $\left|\vec{\Omega}\right|$ is the magnitude of rotation per unit time. For small rotations,

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

The cross product of two vectors may be rewritten as the product of a skew-symmetric matrix and a vector. Applying this to the cross-product $\vec{X} \times \vec{\Omega}$ above, we obtain:

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

We can now rearrange (7) into the convenient matrix form

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

where $\phi = [\ \vec{T}^T \quad \vec{\Omega}^T\ ]^T$ is our motion parameter vector, and where

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

Substitution of the right side of (8) for $\vec{V}$ in (6) produces a single linear equation relating image intensity derivatives to rigid body motion parameters under perspective projection at a single pixel:

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

Much of the previous work on motion and pose estimation from intensity data has used this constraint and variations on it. However, in most of that work, the depth values which appear in the equation are not known, and one must use non-linear estimation techniques to solve for the motion (see [10] for examples). Alternatively, the estimation can be reduced to a linear system through the use of a generic shape model of the object being tracked [2, 3]. By using depth measurements directly in our linear constraint equation, we are able to avoid the non-linear computations required by the former class of approaches, as well as reduce the object shape errors inherent in the latter class of approaches.

3.2 Adding the Depth Constraint

Assuming that video-rate depth information is available to us, we can relate changes in the depth image over time to rigid body motion in a manner similar to that shown for intensity information above. For rigid objects, an object point which appears at a particular image location (x, y) at time t will appear at location ($x + v_x$, $y + v_y$) at time $t + 1$. The depth values at these corresponding locations in image space and time should therefore be the same, except for any depth translation that the object point undergoes between the two frames. This can be expressed in a form similar to (1):

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

The same series of steps described above for deriving the BCCE on rigid body motion can now be used to derive an analogous linear depth change constraint equation (DCCE) on rigid body motion. First-order Taylor series expansion, followed by rearrangement into matrix form, produces

$$-Z_t = \begin{bmatrix} Z_x & Z_y \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} - V_z \tag{11}$$

Use of perspective camera projection to relate image velocities to 3-D world velocities yields

$$\begin{aligned} -Z_t &= \frac{1}{Z}\begin{bmatrix} Z_x & Z_y \end{bmatrix} \begin{bmatrix} f & 0 & -x \\ 0 & f & -y \end{bmatrix} \vec{V} - V_z \qquad (12) \\ &= \frac{1}{Z}\begin{bmatrix} fZ_x & fZ_y & -(Z + xZ_x + yZ_y) \end{bmatrix} \vec{V} \end{aligned}$$

Finally, we again constrain 3-D world velocities to rigid body motion by introducing the $\mathbf{Q}$ matrix

$$-Z_t = \frac{1}{Z}\begin{bmatrix} fZ_x & fZ_y & -(Z + xZ_x + yZ_y) \end{bmatrix} \mathbf{Q}\vec{\phi} \tag{13}$$

This linear equation for relating depth gradient measurements to rigid body motion parameters at a single pixel is the depth analog to equation (9). Note that, in contrast to the BCCE, whose derivation begins with an assumption (see equation (1)) that is an approximation at best, the DCCE is based on the linearization of a generic description of motion in 3-D, and we might therefore expect it to lead to more accurate estimation of motion.

3.3 Orthographic Approximation

In many applications, we can approximate the camera projection model as orthographic instead of perspective without introducing significant error in 3-D world coordinate estimation. For the pose tracking algorithms discussed in this paper, use of orthographic projection greatly simplifies the constraint equations derived in the previous sections, thereby making the solution of linear systems of these equations much less computationally intensive.

Derivation of the orthographic analogs of equations (9) and (13) is straightforward. We replace the perspective projection relationship with the orthographic projection equations $x = X$ and $y = Y$, which in turn imply that $v_x = V_x$ and $v_y = V_y$. Hence, equation (5) is replaced by the much simpler equation

$$\vec{v} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \vec{V} \tag{14}$$

Proceeding through the remainder of the derivations of equations (9) and (13) yields their orthographic projection analogs:

$$-I_t = \begin{bmatrix} I_x & I_y & 0 \end{bmatrix} \mathbf{Q}\vec{\phi} \tag{15}$$

$$-Z_t = \begin{bmatrix} Z_x & Z_y & -1 \end{bmatrix} \mathbf{Q}\vec{\phi} \tag{16}$$

3.4 Integration Over Space and Time

We can write intensity and depth constraint equations of the form of equations (9) and (13) for each pixel location that pertains to the object of interest. Because the intensity constraint equations (9) are linear, we can combine them across N pixels by stacking the equations in matrix form: $\vec{b_I} = \mathbf{H_I}\vec{\phi}$. $\mathbf{H_I} \in \Re^{Nx6}$, where each row is the vector obtained by multiplying out the right side of equation (9) at a single pixel i. $\vec{b_I} \in \Re^{Nx1}$, where the ith element is $-I_t$ at pixel i. The I subscripts on $\mathbf{H_I}$ and $\vec{b_I}$ indicate that they only use the intensity constraints. We can collect the depth constraint equations (13) into an analogous linear system: $\vec{b_D} = \mathbf{H_D}\vec{\phi_D}$. Provided that $N > 6$, the least-squares method may be used to solve either of these systems independently for the motion parameters $\vec{\phi}$.

The intensity and depth linear systems and may also be combined into a single linear system for constraining the motion parameters:

$$\vec{b} = \mathbf{H}\vec{\phi},\text{, where } \mathbf{H} = \begin{bmatrix} \mathbf{H_I} \\ \lambda\mathbf{H_D} \end{bmatrix}, \vec{b} = \begin{bmatrix} \vec{b_I} \\ \lambda\vec{b_D} \end{bmatrix} \tag{17}$$

The scaling factor, λ, controls the weighting of the depth constraints relative to the intensity constraints. When one expects depth to be more reliable than intensity, such as under fast changing lighting conditions, one might want to set λ to a value higher than 1, but under other conditions, such as when depth information is much noisier than intensity, one might prefer to use lower λ values. The least-squares solution to the above equation is

$$\vec{\phi} = -(\mathbf{H^T H})^{-1}\mathbf{H}^T\vec{b} \tag{18}$$

The image pixel data used to build the above linear system are taken only from a restricted region of image support. This support region is the set of pixel locations which we believe correspond to the object, and for which intensity, depth, and their derivatives are well-defined.

The motions estimated between pairs of consecutive frames are simply added together to form an estimate of cumulative object motion over time. It is beneficial to supplement this tracking algorithm with a parallel scheme for deciding when the accumulated error has become substantial, and to reinitialize the object pose estimate at these times.

3.5 Shift of World Coordinate System

We improve the numerical stability of the least-squares solution by translating all of the 3-D spatial coordinates to the centroid $\vec{X_o} = \begin{bmatrix} X_o & Y_o & Z_o \end{bmatrix}^T$ of the supported samples. This transformation affects only the matrix $\mathbf{Q}$, and the motion parameter vector $\vec{\phi}$ will compensate for this change. That is, we can rewrite equation (9) as

$$-I_t = \frac{1}{Z}\begin{bmatrix} fI_x & fI_y & -(xI_x + yI_y) \end{bmatrix} \mathbf{Q}'\vec{\phi}' \tag{19}$$

Figure 1. Synthetic sequence input. First and second images: intensity and depth images for initial frame of both synthetic sequences. Third, fourth, and fifth images: extrema of rotations for first synthetic sequence. Sixth image: extreme of translation for second synthetic sequence.

where $\vec{\phi}' = [\ \vec{T}'^T \quad \vec{\Omega}'^T\ ]^T$, and

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

Equation (13) can be modified similarly. We combine these shifted intensity and depth equations into a single linear system and solve for the motion parameters $\vec{\phi}'$ by least squares, as described above. These motion parameters are in the coordinate system of the object centroid; we would like to transform them back to motion parameters $\vec{\phi}$ in the camera coordinate system. The 3-D velocities $\vec{V}'$ in the shifted coordinate system are described by $\vec{V}' = \vec{T}' - (\vec{X} - \vec{X}_o) \times \vec{\Omega}'$. Since these velocities must also equal the velocities $\vec{V}$ in the camera coordinate system, given by equation (7), it is straightforward to show that

$$\vec{\Omega} = \vec{\Omega}' \text{ and } \vec{T} = \vec{T}' + \vec{X}_o \times \vec{\Omega}' \tag{20}$$

4 Results

4.1 Synthetic Image Sequences

Synthetic image sequences provide ground truth that allow us to quantitatively analyze our technique. We generated synthetic sequences from a detailed polygonal model of the frontal half of a human head. The range and color data used to construct the model were obtained with a rotating Cyberware laser scanner equipped to record registered color and depth data at each point. Color test image sequences were constructed via a standard graphical rendering package, employing a perspective camera projection model. Spatial and time derivatives in intensity and depth were computed using difference filters with small support. Sample intensity and depth images of the model are shown in the leftmost two panels of Figure 1.

For each of the two image sequences discussed in this section, motion between all pairs of successive frames were computed using 1) the BCCE only (with measured depth), 2) the DCCE only, and 3) both constraints together. In addition, for each of these cases, parameters were computed using both the perspective and orthographic versions of the constraints. When combining the intensity and depth constraints into a single linear system, we chose the depth constraint weighting factor, λ, to be the ratio of the mean magnitudes of the I_t and Z_t values. This helps to equalize the contributions of the two sets of constraints toward the least-squares solution.

To evaluate the utility of measured depth and the DCCE, we also implemented a motion estimation system which uses only the BCCE and a simple, generic depth model, as in the class of methods which include [2, 3, 4]. The particular form of generic depth model that we used was a plane parallel to the camera image plane, initialized to be at the depth of the object being tracked. We used this system to compute a fourth set of motion parameters for all of the synthetic sequences, according to both the orthographic and perspective versions of the BCCE.

For each pair of frames, the image support region for computation of motion parameters was taken to be, as a first pass, the intersection of the sets of pixels in each frame for which depth is non-background and for which all spatial derivatives do not include differences with background pixels. However, because sampling and object self-occlusion (e.g. of the neck by the chin, in our sequences) creates large depth gradients which do not remain consistent with object pose during motion, we found it helpful to eliminate from the support map all pixels for which the magnitude of the depth gradients exceeded the mean magnitude by more than several standard deviations.

The first synthetic sequence begins with the face oriented toward the camera. The face then makes a 40 degree rotation downward about the X-axis over the course of 35 frames ($\approx$ 1 degree per frame), and returns to the starting position via the opposite rotation in the next 35 frames. The next 70 frames consist of the same rotation about the Y-axis, and the final 70 frames contain the same rotation about the Z-axis. The first two panels of Figure 1 show the intensity and depth images of the starting position for the sequence, while the third, fourth, and fifth panels show intensity images of the sequence's three extrema of rotation.

Figure 2 shows the three computed rotational pose parameters plotted against time over the course of the sequence, using each of the four methods described above, according to the perspective forms of the constraint equations. All rotational parameters are expressed in terms of Euler angles. The ground truth for the parameters, shown as solid lines in each graph, is the same for each graph: the leftmost solid triangle represents the steady rise in the X-axis rotational parameter from zero to 40 degrees and back to zero, the middle triangle represents the identical sequence of changes in the Y-axis rotational parameter, and the rightmost triangle represents these same changes in the Z-axis parameter. Only one Euler angle should be non-zero at any given time.

The results obtained using only the BCCE with the generic shape model indicate that this method does not perform well for out-of-plane rotations (i.e. rotations about the X-axis and Y-axis). All three Euler angles are non-zero throughout these rotations, and the translational parameters (not shown) were also very inaccurate. The second graph, on the other hand, shows that simply adding measured depth to the BCCE greatly improves the pose estimation. The third graph shows that using the DCCE instead of the BCCE improves the estimation even further. The fourth graph shows that the best results of all are obtained by using the BCCE and DCCE together. The accumulated error at the end of the sequence is quite small despite very large rotations of a rather complex (and incomplete) object.

The second synthetic sequence examines translation in depth, which causes difficulties for many pose and motion algorithms. The artificial face again begins the sequence oriented toward the camera, as shown in the first two panels of Figure 1, then translates steadily and directly away from the camera to the extreme position shown in the rightmost panel of this figure, and finally returns at the same speed to the starting position. The distance between the extreme positions of the face was approximately three times the width of the face model, with the extreme farthest position being about twice as far from the camera as the starting position. The translation between the two extreme positions took place in 150 frames.

Figure 3 shows the three computed translational parameters plotted against time using each of the four methods described above, according to the perspective forms of the constraint equations. The results assume the Z-axis is pointing toward the camera. The ground truth for the parameters, shown as solid lines in each graph, is the same for each graph: both the X- and Y-translation is zero throughout the sequence, while the Z-translation forms a triangle indicating its steady decrease to a position far behind the starting point and its subsequent increase back to the starting point. Again, the results using only the BCCE with a generic depth model indicate that this method does not perform well for translation in depth. Its estimates for the X- and Y-translations are quite noisy, while the Z-translation is greatly under-estimated. As for the first synthetic sequence, the graphs for the BCCE with measured depth and for the DCCE alone show significantly improved results, while the graph for the joint use of the BCCE and DCCE shows the best results of all, with very little accumulated error at the end of the sequence.

In general, orthographic projection results for the two sequences were slightly worse than the perspective results, due to the error in the camera model assumption. Of course, from the orthographic BCCE (15) it is apparent that translation in depth cannot be recovered using only the BCCE, even with accurate depth. We indeed found this shortcoming in our results for the second sequence.

4.2 Real Image Sequence

To test our methods on real data, we recorded an approximately 300 frame (10 second) sequence of registered intensity and depth images using our real-time stereo imaging hardware. The image sequence consists of a person initially facing the camera and then rotating her head toward each of the four image corners in succession. Our goal was to track the motion of the person's head.

As for the synthetic sequences, we computed motion estimates using 1) the BCCE only with a generic shape model, 2) the BCCE only with measured depth, 3) the DCCE only, and 4) the combined BCCE and DCCE with measured depth. We used the perspective forms of the constraint equations throughout. The weighting factor λ was chosen as described for the synthetic sequences. Image support regions were computed automatically by selecting large connected foreground regions with smoothly changing range data. This precludes pixels which have an uncertain depth value, typically due to occlusion or low contrast. Also, unlike in the synthetic sequences, real depth imagery is noisy, and we found it advantageous to smooth the depth images prior to computing gradients.

Figure 4 shows still frame images from the sequence overlaid with graphically rendered axes indicating our pose estimates. The original still frames have been greatly lightened to allow the axes to be seen more easily. For the first frame, shown in the top image in the figure, the axes are rendered according to our camera model so as to appear to be a few inches in front of the person's nose. Two of the axes lie in a plane parallel to the image plane, while the third (the darkest axis) is directed at the camera. We updated the position and orientation of the rendered axes for successive frames according to the recovered motion estimates. Therefore, if our pose tracking algorithm works well, we should expect the axes to continue to appear to be rigidly affixed a few inches in front of the nose as the person moves her head. The middle row of images in the figure shows pose

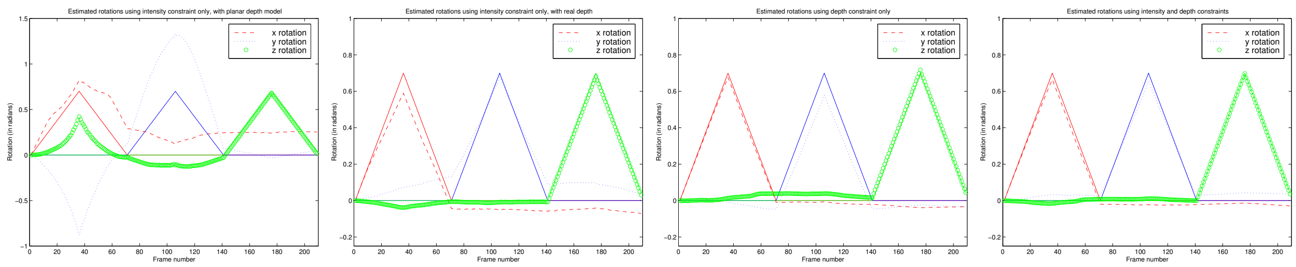

Figure 2. Comparison of ground truth (solid) with computed perspective rotation parameters (see legend) for synthetic rotation sequence, in terms of Euler angles (in radians). Left: BCCE only with planar depth model; Middle-left: BCCE only with real depth; Middle-right: DCCE only; Right: Both constraints used.

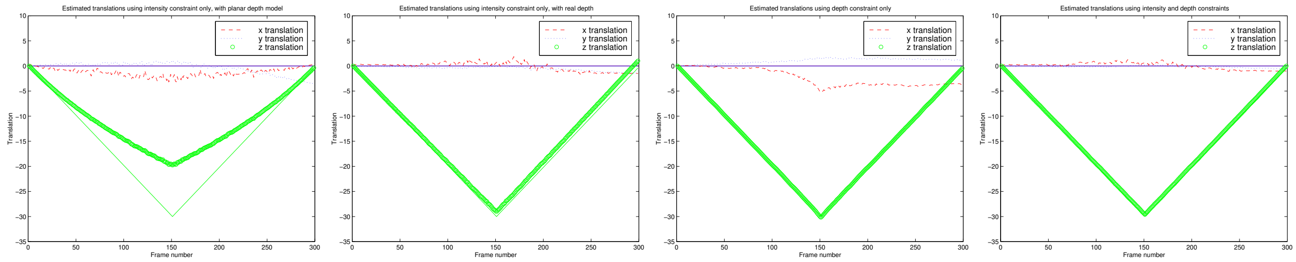

Figure 3. Comparison of ground truth (solid) with computed perspective translation parameters (see legend) for synthetic Z-translation sequence. Left: BCCE only with planar depth model; Middle-left: BCCE only with real depth; Middle-right: DCCE only; Right: Both constraints used.

estimates obtained at several extreme positions in the sequence, using only the BCCE with a generic shape model. The bottom row of images in the figure shows the results obtained for the same frames using the BCCE and DCCE together with measured depth.

The estimates in the bottom row appear to be qualitatively correct in all frames. The final frame shows that this method accumulated very little error over the course of the 300 frame sequence. The estimation also showed very good stability during non-rigid motions, specifically opening and closing of the mouth. In contrast, the results in the middle row of images show that much greater inaccuracy is obtained by using only the BCCE with a generic shape model. This method was not able to cope with even the moderate out-of-plane rotations exhibited in this sequence, as it produced significant spurious translation and exaggerated rotation. For example, the second and fourth frames in the middle row indicate that the head has rotated by over 90 degrees toward the person's right, while the actual rotation is less than 45 degrees. In addition, the last frame in the middle row, showing axes that are far from their initial frame position, reveals that the method accumulated a large amount of error over the course of the sequence.

Results using either the BCCE or the DCCE alone, with measured depth in each case, were not as good as those obtained using the combined constraints. The DCCE alone performed more poorly, producing qualitatively correct but very noisy estimates. The noise in the estimates is likely a result of the significant noise in the depth images themselves.

The quality of the estimates obtained by using the combination of the BCCE and DCCE is much more easily judged by viewing the movies of the above results, which can be found at http://www.interval.com/papers/1999-006. This site also provides result movies for other real and synthetic sequences, as well as a color version of this paper (which allows most of its figures and graphs to be understood more easily).

5 Discussion and Conclusions

We have demonstrated the ability of our method to accurately measure classes of motion, such as translation in depth, rotation out of the image plane, and large angle rotation, which have caused difficulties for most previously described techniques. We have successfully applied the technique to real imagery, and have shown that in practice, the linear brightness and depth constraints complement each other. Finally, we were able to track motion with very little cumulative error over long video sequences (up to 10 seconds, or 300 frames).

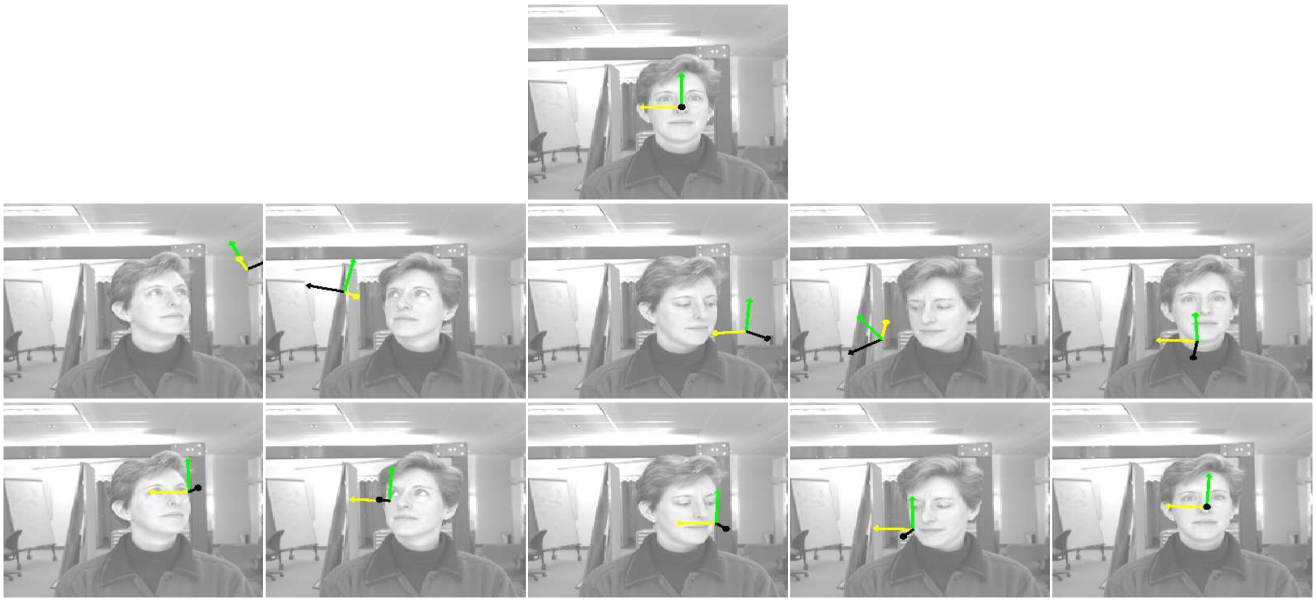

Figure 4. Pose estimation results for real sequence of approx 300 frames (about 10 seconds). Top image: Initial frame in sequence, with result pose axes in initial position. Middle image row: Axes indicate results obtained using BCCE with planar depth for select frames later in sequence. Bottom image row: Results for the same frames using joint BCCE and DCCE with measured depth.

One conclusion that we can draw from the above experiments is that the use of the combined intensity and depth constraints outperforms the use of either independently. This occurs not only because the two constraints together provide more data upon which to base estimates, but also because each type of constraint helps offset the shortcomings of the other. For example, the DCCE is relatively insensitive to photometric effects, while the BCCE typically deals with data that is less noisy and has fewer undefined regions. Another conclusion we can reach is that the substitution of measured, frame-rate depth for generic object shape models can substantially improve on pose estimation methods which use only the BCCE or variations on it. The same is likely true for methods which estimate object shape and motion together.

Our method lends itself nicely to real-time systems, in that it is a linear method which may be solved efficiently via least-squares. Another advantage of our method is that, because it is a differential tracker that updates its shape model of the object over time, it can track object pose through dramatic changes such as those shown in the synthetic head rotation sequence. In fact, we should expect the method to perform reasonably well in tracking an object of arbitrary shape through full 360 degree rotations, assuming small inter-frame motions.

References

[1] J. Bergen , P. Anandan, K. Hanna, R. Higorani. "Hierarchical model-based motion estimation", European Conference on Computer Vision, pp 237-252, 1992.

[2] S. Basu, I. Essa, A. Pentland, "Motion regularization for model-based head tracking", International Conference on Pattern Recognition, 1996.

[3] M. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion". International Conference on Computer Vision, 1995

[4] C. Bregler and J. Malik, "Tracking People with Twists and Exponential Maps", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Santa Barbara, CA), June 1998.

[5] T. Darrell, G. Gordon, M. Harville, J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Santa Barbara, CA), pp. 601-608, June 1998.

[6] G. Gordon, T. Darrell, M. Harville, J. Woodfill, "Background Estimation and Removal Based on Range and Color," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Fort Collins, CO), pp. 459-464, June 1999.

[7] B.K.P. Horn and E.J. Weldon, "Direct Methods for recovering Motion", *International Journal of Computer Vision*, 2:51-76, 1988

[8] J. Shieh, H. Zhuang, R Sudhakar, "Motion Esimtation from a Sequence of Stereo Images: A Direct Method", *IEEE Systems Man and Cybernetics*, SMC(24), No. 7, July 1994, pp. 1044-1053.

[9] G. Stein and A. Shashua, "Direct Estimation of Motion and Extended Scene Structure from a Moving Stereo Rig", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Santa Barbara, CA), June 1998.

[10] T.Y. Tian, C. Tomasi, D. Heeger, "Comparison of Approaches to Egomotion Computation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (San Francisco, CA), pp. 315-320, June 1996.

[11] Woodfill, J., and Von Herzen, B., "Real-Time Stereo Vision on the PARTS Reconfigurable Computer", *Proceedings IEEE Symposium on Field-Programmable Custom Computing Machines*, Napa, pp. 242-250, April 1997.

[12] Zabih, R., and Woodfill, J., "Non-parametric Local Transforms for Computing Visual Correspondence", *Proceedings of the third European Conference on Computer Vision*, Stockholm, pp. 151 - 158. May 1994.