Recursive Motion Estimation of Range Image

Hamid Gharavi *and Shaoshuai Gao*

National Institute of Standards and Technology, USA
Email: {gharavi, sgao}@nist.gov

ABSTRACT

In this paper, we present an innovative recursive motion estimation technique that can take advantage of the in-depth resolution (range) to perform an accurate estimation of objects that have undergone 3-D translational and rotational movements. This approach iteratively aims at minimizing the error between the object in the current frame and its compensated object using estimated motion displacement from the previous range measurements. In addition, in order to use the range data on the non-rectangular grid in the Cartesian coordinate, we consider a combination of derivative filters and the transformation between the Cartesian coordinates and the sensor-centered coordinates. For sequences of moving range images we demonstrate the effectiveness of the proposed scheme.

Index Terms— 3-D motion estimation, range image, object tracking, Ladar, Laser scanners

1. INTRODUCTION

Classical motion estimation techniques in computer vision use intensity images or stereovision to estimate 3-D motion parameters. These techniques are not yet sufficiently robust to be used for highly sensitive real time systems. Recently, with the rapid progress of high-speed range camera technology, capturing what is referred to as 2.5-D images is becoming possible. These images, which provide precise measurements of geometry of the 3-D environment, can make motion estimation and object tracking much easier and more reliable. In general, there are two classes of motion estimation algorithms for range images. Class one is for rigid motion surfaces [1]-[8], and the other is for moving deformable surfaces [9], [10]. Class one can be further divided into two categories. The first is a feature-based algorithm [3], [4], whose performance depends on the detection of reliable range image features and the establishment of interframe correspondence among them. The other is a direct area-based algorithm [1], [2], [5]-[8], which is more straightforward than the feature-based algorithm. In our approach, which falls in this category, we are mainly concerned with rigid motion where the structure of moving images is based on a single beam laser scanner technology. In this technology a deflection mirror assembly scans a beam over the scene. This type of technique has been widely used for many tactical and industrial applications and uses different types of range measurement technologies. One example is the Time of Flight (Pulsed) laser range modules, which send short pulses that are reflected by surrounding objects. Note that with this technology, a three-dimensional scan of a scene is obtained by deflecting the laser beam in equal increments of angle in horizontal and vertical planes. A scanned scene can then be represented in terms of range ρ, horizontal angle θ, and elevation angle ϕ, which corresponds to a spherical (polar) coordinate system.

By converting a range image from the spherical coordinate system to a so-called Cartesian Elevation Map (CEM), Horn and Harris [1] developed a recovery system for the six degrees of freedom of motion of a vehicle, which has been a challenging problem in autonomous navigation. In CEM the depth Z is expressed as a function of X and Y, which corresponds to displacements in the horizontal plane. This time varying CEM is used to estimate translational and rotational movements of rigid objects.

Although the optimized solution offered by Horn and Harris has been very effective, it does not always produce very accurate estimation of 3-D motion displacements, which is crucial for highly sensitive robotic operations. Thus, here we present a recursive approach to enhance estimation accuracy. As will be described next, this iterative approach is based on minimizing the error between the new position of the object and its previous location, after being compensated using estimated motion displacements. In addition, since a set of 3-D points obtained in the CEM coordinate may not be placed regularly on a rectangular grid, we present a method that uses a non-rectangular grid to reconstruct the displaced frame. This scheme employs derivative filters together with transformation between the Cartesian coordinates and sensor-centered coordinates for image reconstruction.

2. 3-D RIGID MOTION ESTIMATION

Recovery of the six degrees of freedom of motion displacement can be best accomplished by using time varying CEM, as proposed by Horn and Harris [1]. Their algorithm is based on the assumption that most of the surface is smooth so that local tangent planes can be constructed. In addition, the motion between frames is smaller than the size of most features in the range image Furthermore, the environment is a single rigid assemblage

I - 673

Report Documentation Page		

1. REPORT DATE **APR 2007**	2. REPORT TYPE	3. DATES COVERED **00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE **Recursive Motion Estimation of Range Image**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Institute of Standards and Technology,100 Bureau Drive, Stop 1070,Gaithersburg,MD,20899-1070**	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002013. Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Held in Honolulu, Hawaii on April 15-20, 2007. Government or Federal Rights

14. ABSTRACT
In this paper, we present an innovative recursive motion estimation technique that can take advantage of the in-depth resolution (range) to perform an accurate estimation of objects that have undergone 3-D translational and rotational movements. This approach iteratively aims at minimizing the error between the object in the current frame and its compensated object using estimated motion displacement from the previous range measurements. In addition, in order to use the range data on the non-rectangular grid in the Cartesian coordinate, we consider a combination of derivative filters and the transformation between the Cartesian coordinates and the sensor-centered coordinates. For sequences of moving range images we demonstrate the effectiveness of the proposed scheme.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**4**	

and only the motion of the sensor relative to the environment has to be recovered.

A time varying CEM can be expressed as $Z(X, Y, t)$; where t denotes time, Z is the depth, and X and Y are displacements in the horizontal and vertical plane, respectively. For a rigid motion scene, the motion can be described as instantaneous translational velocity and instantaneous angular velocity. For every 3-D point, an elevation rate constraint equation relating derivatives of X, Y, Z can be obtained as [1],

$$\dot{Z} = p\,\dot{X} + q\,\dot{Y} + Z_t, \tag{1}$$

where $p = \partial Z/\partial X$, $q = \partial Z/\partial Y$, $Z_t = \partial Z/\partial t$, $\dot{X} = dX/dt$, $\dot{Y} = dY/dt$, $\dot{Z} = dZ/dt$.

The vector to a point on the surface: $\vec{R} = (X, Y, Z)^T$

$$d\vec{R}/dt = -\vec{t} - \vec{\omega} \times \vec{R}, \tag{2}$$

where $\vec{t} = [U \ \ V \ \ W]^T$ is translational velocity and $\vec{\omega} = [A \ \ B \ \ C]^T$ is rotational velocity. Then it has:

$$\begin{cases} \dot{X} = -U - BZ + CY \\ \dot{Y} = -V - CX + AZ \\ \dot{Z} = -W - AY + BX \end{cases}. \tag{3}$$

From (1) and (3),

$$pU + qV - W + rA + sB + tC = Z_t, \tag{4}$$

where $r = -Y - qZ$, $s = X + pZ$, $t = qX - pY$.

Let's assume that there is a set of m pixels in the image and for each such pixel we define the following set of six dimensional vectors for the nth pixel,

$\Phi_n = [p_n \ \ q_n \ \ -1 \ \ r_n \ \ s_n \ \ t_n]^T$, $D = [U \ \ V \ \ W \ \ A \ \ B \ \ C]^T$. From (4) the rate of change for elevation (Z_t) at pixel n can be shown as,

$$(Z_t)_n = \Phi_n^T D. \tag{5}$$

Based on the above equation we can estimate the motion iteratively, where at each iteration the previous estimate is used in the process. Let's assume that in this process two consecutive video frames (generated at a fixed frame rate) are used to measure the change of rate of elevation. After each iteration the estimated motion vectors are used to reconstruct the compensated first frame for the next iteration.

From (5) we can show,

$$noise = (\gamma_n)^{i-1} - (\Phi_n^T)^{i-1}(D - \overline{D}^{i-1}) \tag{6}$$

where γ is the measurement of the displaced frame difference (DFD) between the second frame and the compensated first frame (i.e. the estimated second frame) using the estimated motion vectors [13].

For a cluster of m moving pels, after carrying out the minimization, the least-squares estimate of D is,

$$\sum_{n=1}^{m}(\gamma_n)^{i-1}(\Phi_n)^{i-1} = (\overline{D}^i - \overline{D}^{i-1})\left[\sum_{n=1}^{m}(\Phi_n)^{i-1}(\Phi_n^T)^{i-1}\right]. \tag{7}$$

$$\overline{D}^i = \overline{D}^{i-1} + \left[\sum_{n=1}^{m}(\Phi_n)^{i-1}(\Phi_n^T)^{i-1}\right]^{-1}\sum_{n=1}^{m}(\gamma_n)^{i-1}(\Phi_n)^{i-1}. \tag{8}$$

In order to obtain the new position of each displaced pixel on a non-rectangular grid in CEM, we developed a combination of derivative filters [12] and transformation between the Cartesian coordinates and the sensor-centered coordinates in a non-rectangular grid coordinate.

To use the range data on the non-rectangular sensor grid directly for motion estimation, a new version of the range flow constraint equation is derived in [12]. The three components of the motion vector for one point (i.e. on X, Y, Z directions) can be written as:

$$\begin{cases} \dot{X} = X_x\,\dot{x} + X_y\,\dot{y} + X_t \\ \dot{Y} = Y_x\,\dot{x} + Y_y\,\dot{y} + Y_t \\ \dot{Z} = Z_x\,\dot{x} + Z_y\,\dot{y} + Z_t \end{cases}, \tag{9}$$

where $X_x = \partial X/\partial x$, $X_y = \partial X/\partial y$, $X_t = \partial X/\partial t$, $Y_x = \partial Y/\partial x$, $Y_y = \partial Y/\partial y$, $Y_t = \partial Y/\partial t$, $Z_x = \partial Z/\partial x$, $Z_y = \partial Z/\partial y$, $Z_t = \partial Z/\partial t$, $\dot{x} = dx/dt$, $\dot{y} = dy/dt$, x, y are the sensor grid (range image) index.

Eliminating $\dot{x}$ and $\dot{y}$ then compared with equation (1).

$$\begin{cases} p = (Y_y Z_x - Y_x Z_y)/(X_x Y_y - X_y Y_x) \\ q = (X_x Z_y - X_y Z_x)/(X_x Y_y - X_y Y_x) \\ Z_t = (X_x Y_y Z_t + X_y Y_t Z_x + X_t Y_x Z_y - X_x Y_t Z_y - X_y Y_x Z_t - X_t Y_y Z_x)/(X_x Y_y - X_y Y_x) \end{cases}. \tag{10}$$

In order to reconstruct the first frame after each iteration in a non-rectangular grid, we perform motion compensation directly on the spherical (polar) coordinate. This requires the transformation between (ρ, θ, ϕ) and (X, Y, Z) each time the motion vector estimation is updated. The transformation from sensor-centered coordinates (ρ, θ, ϕ) to Cartesian coordinates (X, Y, Z) can be shown as,

$$\begin{cases} X = \rho \sin\theta \cos\phi \\ Y = \rho \sin\phi \\ Z = \rho \cos\theta \cos\phi \end{cases}. \tag{11}$$

Similarly, from (X, Y, Z) to (ρ, θ, ϕ).

$$\begin{cases} \rho = \sqrt{X^2 + Y^2 + Z^2} \\ \theta = \arctan(X/Z) \\ \phi = \arctan(Y/\sqrt{X^2 + Z^2}) \end{cases}. \tag{12}$$

Given the first frame F_1, and the estimated motion vector MV, the estimated second frame $\hat{F}_2$ will be:

$$\begin{cases} \hat{X}_2(x', y') = X_1(x, y) + MV_X \\ \hat{Y}_2(x', y') = Y_1(x, y) + MV_Y \\ \hat{Z}_2(x', y') = Z_1(x, y) + MV_Z \end{cases}, \tag{13}$$

where x, y, x', y' are the image index. For range data on the rectangular grid, we can directly obtain (x', y') as,

$$\begin{cases} x' = x + MV_X/\Delta X \\ y' = y + MV_Y/\Delta Y \end{cases}, \tag{14}$$

where $\Delta X = X(x+1,y) - X(x,y)$, $\Delta Y = Y(x,y+1) - Y(x,y)$. However, since the 3-D range data in the X, Y, and Z coordinate system are not on the rectangular grid, we cannot directly incorporate the motion vector to reconstruct the motion compensated frame. At the same time, the 3-D points in the sensor centered coordinate (ρ, θ, ϕ) system has the property that $\Delta\theta$ and $\Delta\phi$ are constant, where $\Delta\theta = \theta(x+1,y) - \theta(x,y)$ and $\Delta\phi = \phi(x,y+1) - \phi(x,y)$. Therefore, each time the motion vector is estimated in the X, Y, Z coordinates, motion compensation is performed on the spherical coordinate where,

$$(X_1,Y_1,Z_1) \rightarrow (\rho_1,\theta_1,\phi_1), \ (\hat{X}_2,\hat{Y}_2,\hat{Z}_2) \rightarrow (\hat{\rho}_2,\hat{\theta}_2,\hat{\phi}_2).$$

Then we can obtain (x',y') as,

$$\begin{cases} x' = x + (\hat{\theta}_2 - \theta_1)/\Delta\theta \\ y' = y + (\hat{\phi}_2 - \phi_1)/\Delta\phi \end{cases} \quad (15)$$

3. RESULTS

In order to quantitatively analyze our proposed 3-D motion estimation algorithm we have synthetically generated sequences of moving range images. In particular, these moving images are produced in such a way that a 3-D object can be displaced in accordance with the predefined motion displacement parameters. These images can allow us to evaluate the accuracy of estimated motion vectors with reference to the actual displacement parameters.

Moving range image sequences were constructed via 3-D OOGL (Object Oriented Graphics Library) files. OOGL is a 3-D object data file in which an object is defined by vertices, lines and surfaces. Fig. 1 shows an OOGL file called as "igea", which was selected here to generate a range video sequence for our simulation.

RIF file is a range image format, which is based on the Cartesian coordinates (X, Y, Z components) and consists of the object points and the Mask map (indicates where there are object points). In this format frames with moving objects are constructed by first displacing the object in the OOGL file and then transforming it to the RIF format. In this way we can create a sequence of moving range images (frames) where the object in each frame can be displaced by a predefined 3-D motion vector. In order to assess the performance of the motion estimation, we deliberately corrupted the second range image with zero mean, additive Gaussian noise. Different levels of noise, as described by the standard deviation, are added to the range component, ρ, in the spherical coordinate (before transformation to the CEM coordinate).

Now we present the simulation results of the proposed motion estimation technique in accordance with equation (8). From this equation we can observe that for $i = 1$ (first iteration) and for the initial estimate $D^0 = 0$, (8) reduces to the Horn and Harris algorithm [1]. Therefore, any improvement after the first iteration is credited to the proposed recursive method over the Horn and Harris algorithm. Another factor affecting the performance of the

estimation method is dealing with the non-rectangular grid typical of range images in the X, Y, and Z coordinate system. As described in Section II, we have developed a method which is a combination of the derivative filter and transformation between (ρ, θ, ϕ) and (X, Y, Z).

We use two criteria as a measure of performance: Mean Square Error (MSE) and Motion Vector Error (MVE). The MSE between Frame 1 and Frame 2 is defined as,

$$MSE = \frac{1}{m}\sum_R [(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2],$$

where R is the region that combines both objects in two frames, $R = MASK_1 \bigcup MASK_2$, m is the number of the points in region R.

Given the true motion parameters (U,V,W,A,B,C) and the estimated ones $(\hat{U},\hat{V},\hat{W},\hat{A},\hat{B},\hat{C})$, the MVE is defined as:

$$MVE = \frac{|U-\hat{U}|+|V-\hat{V}|+|W-\hat{W}|+|A-\hat{A}|+|B-\hat{B}|+|C-\hat{C}|}{|U|+|V|+|W|+|A|+|B|+|C|}.$$

In our experiments we set the maximum number of iterations to 16. However, if the MSE difference between successive iterations is less than a threshold (i.e., 0.1) and the current MSE larger than the previous one, the previous estimation will be selected and the iteration will be stopped.
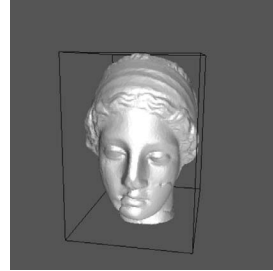

Fig 1. The OOGL files: "igea"

We carried out these experiments under various test conditions. For example, we used different parameters to transform a 3-D image (see Fig. 1) from OOGL to RIF. Based on the 3-D test image shown in Fig. 1, we created a large number of range video sequences with different view angles and different translation and rotational motion.

The results of our experiments are presented subjectively and objectively. In the subjective results we show a difference between the second frame and the estimated second frame. Note that the estimated second frame corresponds to the motion compensated first frame based on the estimated motion parameters (e.g., after each iteration). This frame difference, as shown by equation (6) in Section II, corresponds to the displaced frame difference (DFD). In the objective results, we show the MSE curve and the MVE curve for the recursive motion estimation algorithm. The results, which show two consecutive frames of "igea" image, are depicted in Fig. 2 and 3. It can be clearly observed that the results of the first iteration, which correspond to the Horn and Harris algorithm, are very poor. This is mainly because the surfaces of some objects are not smooth enough and there are many surfaces that are not

always conjoined smoothly. However, after the first iteration, due to the proposed recursive motion estimation algorithm, the estimated motion parameters approach the actual motion parameters.

In order to test the resistance of the motion estimation scheme to noise, we added a different level of synthetic noise (white Gaussian noise) on the second range image. We then averaged the results by running each test 20 times. The results are depicted in Fig. 4. We can see that the performance of the motion estimation drops as the noise level increases. Nevertheless, the recursive motion estimation continues to maintain its gain over the Horn and Harris algorithm.

Finally, we should point out that for every iteration the computational cost would be the same as with the Horn and Harris algorithm, except that additional processing would be required to achieve transformation between (ρ, θ, ϕ) and (*X*, *Y*, *Z*) after each iteration.
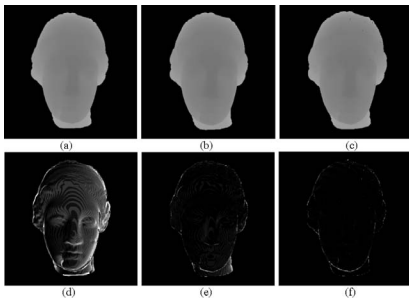

Fig. 2. Subjective evaluations of the proposed motion estimation scheme for "igea". (a) The first image; (b) The second image; (c) The estimated second image using the estimated motion parameters of final iteration; (d) The difference image between the original two images; (e) The difference image between the second image and the estimated second image (DFD) using the motion parameters of the first iteration (Horn and Harris algorithm); (f) The difference image between the second image and the estimated second image (DFD) using the motion parameters of the final iteration.
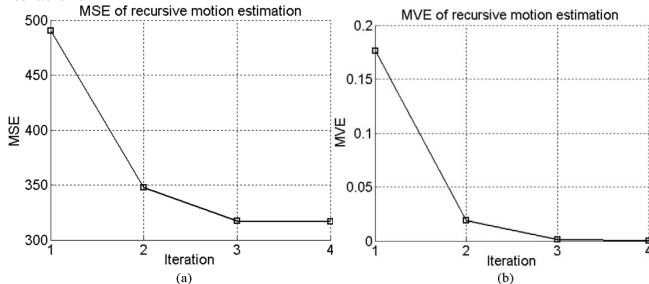

Fig. 3. Objective evaluations of the proposed motion estimation scheme for "igea". (a) MSE; (b) MVE.

4. CONCLUSIONS

In the realm of 3D measurements, high-resolution range moving images that can accurately perform object tracking and velocity estimation would be required for highly sensitive and critical operations. Thus, our main objective has been to improve motion estimation accuracy involving both rotational and translational movements. We have presented a recursive motion estimation technique that can take advantage of the in-depth resolution (range). We have

shown that displacement of objects with complex 3-D motion in range images can be accurately estimated by using the proposed recursive approach. In addition, we presented a method of reconstructing a motion compensated frame in a non-rectangular grid structure typical of range images in the Cartesian coordinate system.
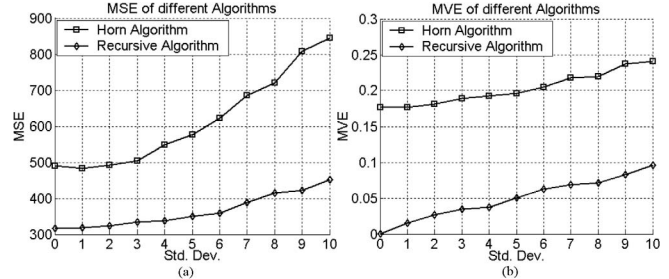

Fig. 4 Objective comparison of different algorithms for "igea" with different level of noise. (a) MSE; (b) MVE.

5. REFERENCES

[1] B. K. P. Horn and J. Harris, "Rigid body motion from range image sequences," *CVGIP: Image Understanding*, vol. 53, no. 1, pp. 1–13, 1991.

[2] K. Chaudhury, R. Mehrotra, and C. Srinivasan, "Detecting 3-D motion field from range image sequences," *IEEE Trans. Syst., Man, and Cybern. B*, vol. 29, pp. 308-314, 1999.

[3] K. Arun, T. Huang, and S. Blostein, "Least square fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 698–700, 1987.

[4] B. Sabata and J. K. Agarwal, "Estimation of motion from a pair of range images," *CVGIP: Image Understanding*, vol. 54, no. 3, pp. 309–324, 1991.

[5] M. Harville, A. Rahimi, T. Darrell, G. Gordon, and J. Woodfill, "3d pose tracking with linear depth and brightness constraints," In *ICCV*, pp. 206-213, 1999.

[6] Y. Liu and M. A. Rodrigues, "Correspondenceless motion estimation from range images," In *ICCV*, pp. 654-659 1999.

[7] L. Luchese, G. Doretto, and G. M. Cortelazzo, "Frequency domain estimation of 3-d rigid motion based on range and intensity data," In *Recent Advances in 3-D Digital Imaging and Modeling*, pp. 107-112, 1997.

[8] R. Szeliski, "Estimating motion from sparse range data without correspondence," In *ICCV*, pp. 207-216, 1988.

[9] M. Yamamoto, P. Boulanger, J. Beraldin, and M. Rioux, "Direct estimation of range flow on deformable shape from a video rate range camera," *IEEE Trans. Pattern Anal. Machine Intell.*, *PAMI-15*, pp. 82-89, 1993.

[10] L. Tsap, D. Goldgof, and S. Sarkar, "Model-based force-driven nonrigid motion recovery from sequences of range images without point correspondences," *Image and Vision Computing*, vol. 17, pp. 997-1007, 1999.

[11] G. Hetzel, B. Leibe, P. Levi, B. Schiele, "3D Object Recognition from Range Images using Local Feature Histograms, " *Proceedings of CVPR 2001*, vol. 2, pp. 394-399, Kauai Island, Hawaii, Dec. 2001.

[12] H. Spies, "Analysing Dynamic Processes in Range Data Sequences," PhD thesis, University of Heidelberg, 2001.

[13] H. Gharavi, and H. Reza-Alikhani, "Pel-recursive motion estimation algorithm," *Electronics Letters*, vol. 37, pp. 1285-1286, 2001.