ROBUST REAL-TIME EGOMOTION FROM STEREO IMAGES

Louis-Philippe Morency

Artificial Intelligence Laboratory, MIT
Cambridge, MA
lmorency@ai.mit.edu

Rakesh Gupta

Honda Research Institute USA, Inc.
Mountain View, CA
rgupta@hra.com

Abstract

In this paper, we present a novel technique for estimating large camera displacement using stereo images. The relative transformation between two stereo image pairs is estimated using a hybrid registration algorithm which combines the robustness of multi-scale feature tracking for large movements and the accuracy of 3D normal flow constraints. Our hybrid technique takes advantage of depth information available from the stereo camera which makes it less sensitive to lighting variations. We tested the accuracy of our hybrid algorithm on real stereo sequences and showed that our technique handles displacements up to 150 cm and rotations up to 20 degrees between images. Our algorithm runs at 6 Hz on a Pentium 4 1.7GHz.

1. INTRODUCTION

Egomotion estimation is an active research topic in computer vision with many applications like robot localization, merging of aerial imagery, and 3D reconstuction. By using gradient-based techniques on all valid pixels, sub pixel accuracy can be reached for small displacements [4]. When stereo images are taken from widely separated viewpoints, feature-based tracking [1, 2, 3] can estimate relative transformation robustly, but there is no easy way to verify potential matches. New techniques for matching features over large displacement and minimizing error accurately are therefore necessary.

This paper presents our hybrid technique for large displacement egomotion estimation from stereo images. The pose of the stereo camera is estimated using a hybrid registration algorithm which combines the robustness of multi-scale feature tracking for large movements and the accuracy of 3D normal flow constraint. Our hybrid technique takes advantage of the depth information from the stereo camera which makes it less sensitive to lighting variations.

When no depth information is available, multi-scale feature tracking can estimate 2D affine transformation accurately [1]. When depth is available from stereo cameras, this technique can be extended to 3D rigid transformation by matching features using the 2D multi-scale tracker, projecting those 2D features in 3D, and finally minimizing the Euclidian distance between them. This approach is similar to the Iterative Closest Point (ICP) algorithm [5, 6] and the point-to-point error function.

The 3D normal flow constraint [7] minimizes an error distance, which is a function of both appearance and depth information. By using this gradient-based error function, sub pixel accuracy can be reached for small displacements [4]. Because of its simplicity and fast speed, this technique can be applied iteratively on all valid depth pixels to extend its range of displacements and still keep accurate estimation.

In the following section we describe the general framework of our hybrid registration algorithm. In section 3, we describe the multi-scale feature tracking technique and how least-median-square filtering removes outliers. In section 4, we describe the 3D normal flow constraint minimization. In section 5, we describe experimental results on still images with ground truth and dynamic video sequences.

2. HYBRID REGISTRATION

Our hybrid registration algorithm combines the robustness of a multi-scale feature tracking technique with the accuracy of the 3D normal flow constraint. As shown in Figure 1, both registration techniques share the same iterative framework: correspondences search, error minimization, warping and convergence check.

Our registration algorithm takes two image sets as input: the new image set $\{I_t, Z_t\}$ grabbed at time t and the reference image set $\{I_r, Z_r\}$ (see Figure 2). The reference image set can be either the image set grabbed at time t-1, the first image set, or any relevant image set between time 0 and time t-1 [8]. In our case, the referential image is a keyframe with known pose acquired previously.

The new image set $\{I_t, Z_t\}$ is preprocessed in concert with known camera calibration information to obtain the 3D vertex set Ψ_t of $i := 1..m$ vertices $\vec{v}_{ti} = \{\vec{p}_{ti}, I_{ti}\}$ where $\vec{p}_{ti}$ is the 3D point coordinates in the camera reference and I_{ti} is the brightness value of the point $\vec{p}_{ti}$ as specified by the intensity image I_t.

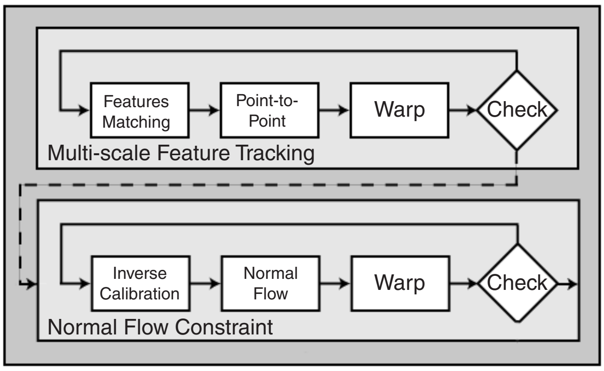


Fig. 1. Flow diagram of our hybrid registration.

The goal of the registration algorithm is to find the rigid pose change $\{\mathbf{R}, \vec{\mathbf{t}}\}$ between the two image sets, where $\mathbf{R}$ is a 3x3 rotation matrix and $\vec{t}$ is a 3D translation vector. At each iteration, a transformation $\vec{\delta}$ represented by 6 parameters vector $[\ \vec{\omega} \quad \vec{t}\ ]^t$ is computed. In this vector, $\vec{\omega}$ is the instantaneous rotation (3 parameters) and $\vec{t}$ is the translation (3 parameters). The current pose estimation is updated as follows:

$$\mathbf{R^{k+1}} = \mathbf{R^k}\mathbf{R}^{(\delta)} \quad (1)$$

$$\vec{t}^{\,k+1} = \vec{t}^{\,k} + \vec{t}^{\,(\delta)} \quad (2)$$

where k is the iteration number and $\mathbf{R}^{(\delta)}$ is the 3x3 matrix representing the rotation $\omega^{(\delta)}$. Initially, $\mathbf{R^0}$ is set to the identity matrix and $\vec{t}^{\,0}$ is set to 0.

The convergence check stage computes the convergence factor ϵ by averaging the distance D between warped 3D points $\vec{p}_{ti}{}'$ and referential 3D points $\vec{q}_{ri}$. If the difference between the convergence factor ϵ of two consecutive iterations is smaller then a threshold value τ, then convergence is reached. The 3D view registration is completed when convergence is reached or, in the case of non-convergence, when a maximum number N_I of iterations is performed.

3. MULTI-SCALE FEATURE TRACKING

Given the referential image I_r, we search for a set of good features to track Ω_r using Shi and Tomasi technique [9]. The following step is the matching step where we search for correspondence between the current image and the referential image. To handle large displacements, we use a pyramidal version of Lucas and Kanade optical flow technique [10]. This algorithm gives us the corresponding set of features Ω_t in the current image.

Using the depth images Z_t and Z_r, we can project the 2D features from feature sets Ω_t and Ω_r in the 3D world. This step gives us two sets of 3D vertices Ψ_t and Ψ_r. To estimate the rigid transformation between both sets, we minimize the point-to-point distance [5] between a vertex $\vec{q}_{ri}$ and the corresponding vertex $\vec{p}_{ti}$:

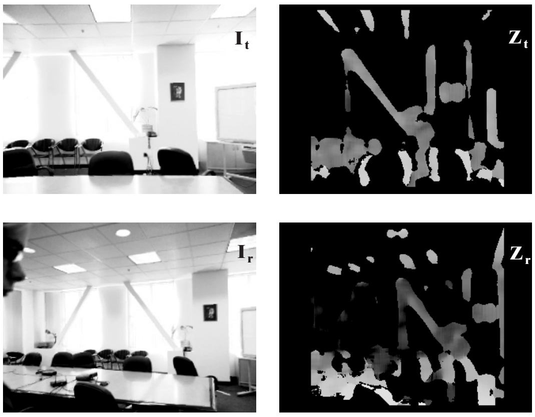


Fig. 2. Example of current and referential image set. In the depth images (right), lighter pixels denote close objects, darker pixels represent further objects and black pixels mean invalid depth.

$$D_{Point}(\vec{q}_{ri}, \vec{p}_{ti}) = (\vec{q}_{ri} - (\mathbf{R}\vec{p}_{ti} - \vec{t})) \quad (3)$$

By approximating the rotation $\mathbf{R}$ with an instantaneous rotation ω and rearranging the equation 3 adequately, we obtain a linear system which defines the error function that we would like to minimize.

To reduce mismatches in the correspondence set Ω_t, we minimize the point-to-point error function using the least-median-square technique of Rousseeuw and Leroy [11]. This technique randomly picks some small subset of the correspondence set, computes an estimate of the pose between frames, and applies this estimate on all remaining points to compute the residual error. The best subset is one that minimizes the median of the residual error. This subset is used to filter the outliers. All the points outside a standard deviation from the median of the best set are filtered out and will not be used during the final least-mean-square minimization.

4. 3D NORMAL FLOW CONSTRAINT

The normal flow constraint is a gradient-based approach which can estimate sub pixel movements accurately. The normal flow constraint is applied on all valid depth pixels. The vertex set of all valid pixels, $\vec{p}_{ti}$, is warped according to the relative pose estimated during the multi-scale feature tracking step. Then, as a matching stage, we use an inverse calibration method to find corresponding points that belong on the same projective ray. This provides the correspondence needed to compute the temporal gradient term of the normal flow constraint.

4.1. Inverse Calibration

The inverse calibration approach searches for corresponding points of $\vec{p}_{ti}$ by projecting vertices from the 3D coordinate system of Ψ_t to the referential depth image Z_r coordinate system:

$$\begin{bmatrix} \vec{u}_{ri} \\ 1 \end{bmatrix} = \mathbf{C} \begin{bmatrix} \vec{p}_{ti} \\ 1 \end{bmatrix} \tag{4}$$

where $\mathbf{C}$ is a 3x4 projection matrix that relate 3D coordinate system of $\vec{p}_{ti}$ to the 2D image coordinate $\vec{u}_{ri} = \begin{bmatrix} u_{ri} & v_{ri} \end{bmatrix}$. This matrix is based on the stereo camera intrinsic parameters.

The 3D coordinates of the referential image set, $\vec{q}_{ri} = \begin{bmatrix} x_{ri} & y_{ri} & z_{ri} \end{bmatrix}$, are interpolated from the depth image Z_r as follows:

$$z_{ri} = Z_r(\vec{u}_{ri}) \quad , \quad x_{ri} = f\frac{u_{ri}}{z_{ri}} \quad , \quad y_{ri} = f\frac{v_{ri}}{z_{ri}} \tag{5}$$

4.2. Normal Flow Constraint

Given 3D input data, the normal flow is the component of the optical flow in the direction of the image gradient. As shown in [7], the normal flow can be expressed as:

$$-\frac{\partial I_{ri}}{\partial t} = \nabla I_{ri} \left[\frac{\partial \vec{u}_{ri}}{\partial \vec{q}_{ri}} \right] \vec{V} \tag{6}$$

where $\nabla I_{ri} = \begin{bmatrix} \frac{\partial I_{ri}}{\partial u_{ri}} & \frac{\partial I_{ri}}{\partial v_{ri}} \end{bmatrix}$ is the image gradient, $\vec{V} = \begin{bmatrix} \frac{\partial x_{ri}}{\partial t} & \frac{\partial y_{ri}}{\partial t} & \frac{\partial z_{ri}}{\partial t} \end{bmatrix}$ is the velocity of the object and $\frac{\partial I_{ri}}{\partial t}$ is the time gradient. $\frac{\partial I_{ri}}{\partial u_{ri}}$ and $\frac{\partial I_{ri}}{\partial v_{ri}}$ are computed directly from the referential image I_r. The time gradient is approximated by:

$$\frac{\partial I_{ri}}{\partial t} = I_{ti} - I_{ri} \tag{7}$$

For a perspective projection where $u_{ri} = f\frac{x_{ri}}{z_{ri}}$ and $v_{ri} = f\frac{y_{ri}}{z_{ri}}$, we can find the Jacobian matrix:

$$\frac{\partial \vec{u}_{ri}}{\partial \vec{q}_{ri}} = \begin{bmatrix} \frac{f}{z_{ri}} & 0 & -f\frac{x_{ri}}{z_{ri}^2} \\ 0 & \frac{f}{z_{ri}} & -f\frac{y_{ri}}{z_{ri}^2} \end{bmatrix} \tag{8}$$

Since the object is rigid, the velocity V can be expressed as:

$$\vec{V} = \begin{bmatrix} \mathbf{I} & -\hat{q}_{ri} \end{bmatrix} \vec{\delta} \tag{9}$$

where $\mathbf{I}$ is a 3x3 identity matrix and $\hat{q}_{ri}$ is the skew matrix of the vector $\vec{q}_{ri}$. By rearranging the equation, we get a linear system:

$$\varepsilon_{NFC} = \|\mathbf{A_{NFC}}\vec{\delta} - \vec{b}_{NFC}\|^2 \tag{10}$$

where each line is defined as follow

$$\vec{A}_i = \nabla I_{ri} \left[\frac{\partial \vec{u}_{ri}}{\partial \vec{q}_{ri}} \right] \begin{bmatrix} \mathbf{I} & -\hat{q}_{ri} \end{bmatrix} \tag{11}$$

$$b_i = -\frac{\partial I_{ri}}{\partial t} \tag{12}$$

5. EXPERIMENTS

We tested our hybrid tracker with sequences obtained from a stereo camera. We used the Small Vision System [12] to calibrate our stereo camera and to compute the disparity image. Without special optimizations, our hybrid registration algorithm can update poses for 320x240 size images at 6Hz on a Pentium 4 1.7GHz.

To show the accuracy of our hybrid technique, we compared three different registration techniques on a dynamic sequence as well as still images. The first technique is the multi-scale feature tracking and point-to-point error function minimization using normal mean-least-square approach. The second technique is a robust version of technique where the point-to-point error function is minimized using the least-median-square approach (described in section 3). The third technique is our hybrid registration algorithm which combines feature tracking, least-median-square and normal flow constraint.

5.1. Dynamic Sequence

For the first experiment, we compared all three techniques on a dynamic sequence where the camera moves mostly straight in the Z direction for 100 cms. The total sequence is 437 frames. We registered each frame to the first frame. The goal of this experiment is to test the robustness of our algorithm when the camera is moving. The final trajectory should be straight along the Z axis. For each registration technique, Figure 3 shows 3 of the 6 degrees of freedom of the trajectory: translation in X, translation in Y and rotation around Y. We can observe in the X translation and Y translation graphs that the hybrid technique gives better and stable results. In the rotation around Y graph, we can see that at approximately frame 250, the estimate shows rotation along the positive direction. This happened because the camera didn't exactly follow a straight trajectory and between frames 250 and 350 the camera was a slightly rotated toward the left.

5.2. Still Images

For the second experiment, we applied our hybrid algorithm on two still images (see Figure 2) taken from the same stereo camera at different location and orientation. The current image set $\{I_t, Z_t\}$ was taken 90±2 cm along the X axis and 100±2 cm along the Y axis away from the referential image

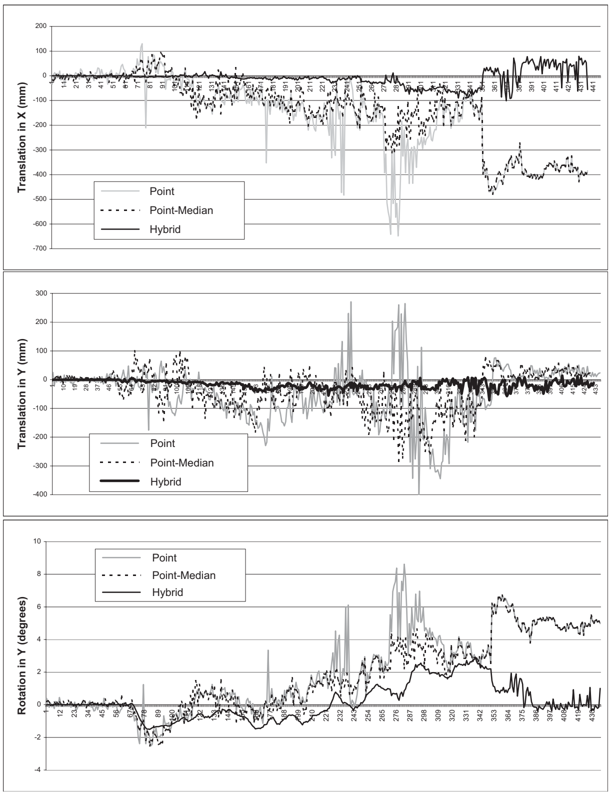
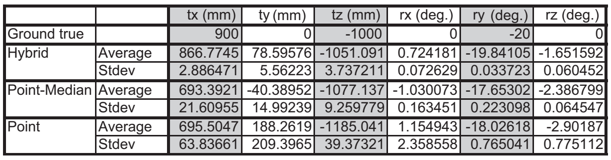


Fig. 3. Results from the dynamic sequence experiment. The horizontal axis represents the frame index.



		tx (mm)	ty (mm)	tz (mm)	rx (deg.)	ry (deg.)	rz (deg.)
Ground true		900	0	-1000	0	-20	0
Hybrid	Average	866.7745	78.59576	-1051.091	0.724181	-19.84105	-1.651592
	Stdev	2.886471	5.56223	3.737211	0.072629	0.033723	0.060452
Point-Median	Average	693.3921	-40.38952	-1077.137	-1.030073	-17.65302	-2.386799
	Stdev	21.60955	14.99239	9.259779	0.163451	0.223098	0.064547
Point	Average	695.5047	188.2619	-1185.041	1.154943	-18.02618	-2.90187
	Stdev	63.83661	209.3965	39.37321	2.358558	0.765041	0.775112

Table 1. Results from the still images experiment.

set. The camera was also rotated around the Y axis 20±1 degrees. The wall was approximately 5 meters away from the referential camera. We repeated the same experiment 10 times.

The registration results from each techniques are presented in table 1. The average and standard deviation are over all 10 runs. The feature-based tracking alone doesn't give good results since outliers are present and not all the available information is used. The robust version of the feature tracking removes outliers and improves the pose estimate. This improvement can be seen in the Y translation estimate and the rotations around X and Z axis. The hybrid registration technique gives the best results.

6. CONCLUSION

We presented an efficient technique to register 3D views with large displacements and rotations. Our registration framework merges the robustness of the multi-scale feature tracking with the accuracy of the normal flow constraint. We tested our technique on different egomotion problems and show an accuracy of 2 cms for displacements as large as 150 cms and rotations up to 20 degrees.

7. REFERENCES

[1] Rakesh Kumar and Allen R. Hanson, "Robust methods for estimating pose and a sensitivity analysis," *CVGIP: Image Understanding*, vol. 60, no. 3, pp. 313–342, 1994.

[2] C. Schmid and R. Mohr, "Local greyvalue invariants for image retrieval," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 19, no. 5, pp. 872–877, 1997.

[3] P. Torr and C. Davidson, "Impsac: Synthesis of importance sampling and random sample consensus," in *In ECCV*, 2000, pp. 819–833.

[4] J.L. Barron, D.J. Fllet, and S.S. Beauchemin, "Performance of optical flow techniques," *IJCV*, vol. 12, no. 1, pp. 43–77, 1994.

[5] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 14, no. 2, pp. 239–256, February 1992.

[6] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," in *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 1991, pp. 2724–2728.

[7] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade, "Three-dimensional scene flow," in *ICCV (2)*, 1999, pp. 722–729.

[8] L.-P. Morency, A. Rahimi, and T. Darrell, "Adaptive view-based appearance models," in *Computer Vision and Pattern Recognition*, 2003.

[9] J. Shi and C. Tomasi, "Good features to track," in *CVPR94*, 1994, pp. 593–600.

[10] B. D. Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision," in *In IJCAI*, 1981, pp. 674–679.

[11] Rousseeuw P.J. and Leroy A.M., *Robust Regression and Outlier Detection*, John Wiley and Sons, 1987.

[12] Videre Design, *MEGA-D Megapixel Digital Stereo Head*, http://www.ai.sri.com/ konolige/svs/, 2000.