ROBUST EGO-MOTION ESTIMATION AND 3D MODEL REFINEMENT USING DEPTH BASED PARALLAX MODEL

Amit K Agrawal and Rama Chellappa

University of Maryland
Department of Electrical and Computer Engineering
College Park, MD 20742 USA

ABSTRACT

We present an iterative algorithm for robustly estimating the ego-motion and refining and updating a coarse, noisy and partial depth map using a *depth based parallax model* and brightness derivatives extracted from an image pair. Given a coarse, noisy and partial depth map acquired by a range-finder or obtained from a Digital Elevation Map (DEM), we first estimate the ego-motion by combining a global ego-motion constraint and a local brightness constancy constraint. Using the estimated camera motion and the available depth map estimate, motion of the 3D points is compensated. We utilize the fact that the resulting surface parallax field is an epipolar field and knowing its direction from the previous motion estimates, estimate its magnitude and use it to refine the depth map estimate. Instead of assuming a smooth parallax field or locally smooth depth models, we locally model the parallax magnitude using the depth map, formulate the problem as a generalized eigen-value analysis and obtain better results. In addition, confidence measures for depth estimates are provided which can be used to remove regions with potentially incorrect (and outliers in) depth estimates for robustly estimating ego-motion in the next iteration. Results on both synthetic and real examples are presented.

1. INTRODUCTION

3D scene reconstruction and ego-motion estimation has been an active area of research over the past few decades. Dynamic scene analysis requires estimation of the relative motion between the camera, scene and the scene structure in the form of a depth map. Motion estimation of a camera moving in an environment is useful for tasks such as navigation, obstacle-detection etc. and recovering the scene structure helps in enhanced visualization and building 3D models of the scene. With increased use of range scanners and DEM's, there is considerable interest in fusing the depth information provided by them with the information from the image sequences to develop robust algorithms for building enhanced 3D models. The available depth information, however, is often noisy, coarse and partial (may lack data at certain regions). In this paper, we address the problem of using such noisy, coarse and partial depth information along with intensity images to estimate the ego-motion and the depth map of the scene.

Several researches have worked on the problem of ego-motion estimation and depth recovery using intensity images. Feature based methods [1][2] use features or tokens to get depth information and motion. Flow based methods [3][4] assume that optical flow is available. Direct methods [5][6][7][8][9][10][11] do not require intermediate steps such as feature extraction or flow computation and work directly with spatial and temporal image gradients. Most of the previous approaches assumes locally smooth depth models for estimating depths [7][10] or small depth variations compared to the distance from the camera [3][8]. However, these assumptions are violated when the depth variations may be large (for example, in urban environments) and at depth boundaries. The effect of noise in available data may require a non-smooth local depth refinement. We show how to use the epipolar constraint and model the parallax field appropriately to deal with such cases. Parallax based approaches proposed in [9][11] assume a dominant planar region to be present in the image or the presence of a small planar region for motion estimation [12]. Our approach does not require any such assumptions. Also, many of the previous methods use the information from the entire image for estimating ego-motion which may not be useful and can even contribute to errors. We show how to discard potentially erroneous image regions that may include incorrect depth estimates and ambiguities arising from the presence of local edge structure in the direction of the focus of expansion by incorporating a confidence measure in estimating depths.

We describe our algorithm in detail in Section 2. Section 3 presents results on both synthetic and real models and comparisons of the recovered depth map and ego-motion using the proposed model with those obtained using a constant parallax model. This is followed by conclusions in section 4.

2. ALGORITHM

Our method is a direct approach that uses two intensity images (referred to as *key* and *offset* frames) and an initial coarse, noisy and partial depth map (referred to as *reference depth map*) to estimate the ego-motion and the depth map in an iterative fashion (we call these iterations *global iterations*). We start with estimating the ego-motion given the reference depth map and refining the available depth map using the estimated ego-motion iteratively, until the motion estimates converge or a specified number of iterations have been reached.

Assuming brightness constancy, we have $I(\mathbf{r},t) = I(\mathbf{r}-\mathbf{u},t-$

Prepared through collaborative participation in the Advanced Decision Architectures Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0009. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The authors would like to thank Phil David and Jeff DeHart of U. S. Army Research Laboratory for helpful discussions.

1) where $I(\mathbf{r},t)$ and $I(\mathbf{r},t-1)$ denote the key and offset frames respectively. Then the 2D image motion $\mathbf{u}$ is given by [6]

$$\mathbf{u} = AhT + B\Omega \tag{1}$$

where $B = \begin{bmatrix} \frac{xy}{f} & -(f+\frac{x^2}{f}) & y \\ (f+\frac{y^2}{f}) & -\frac{xy}{f} & -x \end{bmatrix}, A = \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix}$, $h = \frac{1}{Z}$ and (T,Ω) denote the translational and rotational velocities. For estimating ego-motion and depth, we minimize the deviations from the brightness constancy equation

$$E = \sum_R (I(\mathbf{r},t) - I(\mathbf{r}-\mathbf{u},t-1))^2 \tag{2}$$

over suitable regions R. A way to minimize (2) is to perform iterative Gauss-Newton minimization (we call these iterations *local iterations*). Let $\delta\mathbf{u}$ denote the incremental 2D motion for a local iteration due to motion refinement or depth refinement. The appropriate motion (or depth) refinement can be estimated by minimizing

$$E = \sum_R (\nabla I^T \delta\mathbf{u} + \Delta I)^2 \tag{3}$$

with respect to $\delta\mathbf{u}$, where $\nabla I = [I_x, I_y]^T$ denotes the spatial image derivatives and $\Delta I = I(\mathbf{r},t) - I(\mathbf{r}-\mathbf{u},t-1)$ denotes the difference of the key image and the warped offset image according to current depth and motion estimates. In what follows, we describe the motion estimation and depth refinement steps in detail.

2.1. Ego-Motion estimation given a depth map

Let Z_i denote the current estimate of the depth map (reference depth map or estimated from a previous global iteration) with i denoting the global iteration index. To estimate the ego-motion, we minimize (2) with respect to T and Ω using Z_i as the depth map. The region R is decided on the basis of the confidence measure provided by the depth refinement phase as described in section 2.2 (for the first global iteration we use the entire image region).

Let T_i, Ω_i denote the ego-motion estimate from the previous global iteration (for the first global iteration, we use $T = [0,0,1]^T$, $\Omega = [0,0,0]^T$). Within each global iteration, we refine the ego-motion estimate by performing local iterations as follows. Let $\delta T, \delta\Omega$ be the incremental ego-motion update for a local iteration. Using (1), we have $\delta\mathbf{u} = Ah_i\delta T + B\delta\Omega$ where $h_i = \frac{1}{Z_i}$. $\delta T, \delta\Omega$ can be obtained by minimizing (3) with respect to $\delta T, \delta\Omega$ with $\delta\mathbf{u}$ as above. This is a linear system in $\delta T, \delta\Omega$ and a least square solution is obtained. The local iterations are performed until the error E in (3) stops decreasing.

2.2. Depth refinement using ego-motion

We now show how to refine the depth map given an estimate of the ego-motion and the available depth information. Let T_i, Ω_i denote the current ego-motion estimate and Z_i denote the available depth map estimate. Let δZ be the incremental depth map estimate and $Z = Z_i + \delta Z$ be the refined depth map. Using equation (1) incremental 2D motion can be written as

$$\delta\mathbf{u} = A(h - h_i)T_i \tag{4}$$

where $h = \frac{1}{Z}$. Thus, the incremental motion due to depth refinement (*surface parallax field*) is in the direction of the focus of expansion (FOE), i.e it is an epipolar field. Since we have an estimate of the FOE (defined as (x_f, y_f)) from T_i, for each pixel (x,y) we have

$$\delta\mathbf{u} = \beta\mathbf{du} \tag{5}$$

where $\mathbf{du}(x,y) = [\frac{(x-x_f)}{\sqrt[2]{(x-x_f)^2+(y-y_f)^2}}, \frac{(y-y_f)}{\sqrt[2]{(x-x_f)^2+(y-y_f)^2}}]^T$ denotes the parallax direction and β denotes the parallax magnitude. Using (5), (3) can be written as

$$E = \sum_R (I_p\beta + \Delta I)^2 \tag{6}$$

where $I_p = \nabla I^T \mathbf{du}$ denotes the projection of the intensity gradient on the parallax direction. The region R for depth refinement is chosen to be a local neighborhood of $N \times N$ pixels. We first minimize (6) to get an estimate of β and then use it to obtain Z from (4) and (5). Thus, the ambiguity arising from the aperture problem has been resolved because the incremental 2D motion is constrained to lie along a line passing through the FOE.

For estimation problems such as above, a smoothness constraint is generally applied. For example, in optical flow estimation, it is often assumed that the flow is constant within a neighborhood or is a parametric function [13] that imposes smooth flow. The smoothness constraint on depths can be applied by assuming a smooth depth model (constant or planar) over the neighborhood (as in [7][10]) and directly using (4) and (3) to estimate Z. However, these assumptions are violated at depth boundaries. Also, the effect of noise in available depth map estimate (from a range finder or from the previous iterations) may require a non-smooth depth refinement within the neighborhood. Thus in such cases, the parallax magnitude is *not* smooth over the neighborhood. From (4) and (5), we observe that the parallax magnitude β has a dependence on $\frac{1}{Z_i^2}$. Therefore, we propose to use the following *depth based parallax model (DBPM)*

$$\beta = a_0 + \frac{a_1}{Z_i} + \frac{a_2}{Z_i^2} \tag{7}$$

where the parameters a_0, a_1 and a_2 are assumed to be constant within the neighborhood. Note that even though we use a parametric model, it allows the parallax magnitude to vary non-uniformly within the region since the model is based on depth values that can vary non-uniformly within the region. This in turn, allows discontinuity preserving depth refinement within the region.

We minimize (6) by formulating it as a generalized eigenvalue problem to obtain a total least squares (TLS) solution. Let $\gamma = [\beta_1, \beta_2]^T = Bp$, where $B = \begin{bmatrix} 1 & h_i & h_i^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & h_i & h_i^2 \end{bmatrix}$ and $p = [a_0, a_1, a_2, a_3, a_4, a_5]^T$ denote the parameters to be estimated. The parallax magnitude β will then be given by $\beta = \frac{\beta_1}{\beta_2}$. Equation (6) can be written as $E = \sum_{N\times N} \gamma^T g g^T \gamma$ where $g = [I_p, \Delta I]^T$. To avoid the trivial solution $\gamma = 0$, the constraint $\gamma^T\gamma = 1$ is imposed. Using Lagrange multipliers, the error function can be written as

$$E = p^T \sum_{N\times N} (B^T g g^T B)p + \lambda(1 - p^T B^T B p) \tag{8}$$

Differentiating with respect to p, we get $Gp = \lambda Dp$, where $D = B^T B$ and $G = \sum_{N\times N} B^T g g^T B$. Since the rank of D is two, there will be only two valid generalized eigen-value/eigen-vector pair. Let $\lambda_1 \geq \lambda_2$ be the valid generalized eigen-values. The generalized eigen-vector corresponding to λ_2 will be the solution for p. Consider the following scenarios:

1. Homogeneous regions: No intensity variation in spatio temporal direction. $\lambda_1 = \lambda_2 = 0$
2. Intensity gradient is in direction perpendicular to the parallax direction, i.e. $I_p = 0$. $\lambda_1 > 0, \lambda_2 = 0$
3. Intensity variation in accordance with ΔI. $\lambda_1 > 0, \lambda_2 = 0$
4. Intensity variation in all directions. No sufficient structure. $\lambda_1 > 0, \lambda_2 > 0$

Confidence measures based on eigen-values and/or condition number have been proposed in [13][14]. We use $C = (\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2})^2$ as the confidence measure for depth estimation. Homogeneous regions (case 1) can be identified by using a threshold on the sum of eigen-values. Regions where local edge structure is aligned along the parallax direction (case 2) can be identified by thresholding I_p. For all such regions, C is set to zero. Thus, C is close to one when the parallax magnitude can be estimated reliably (case 3) and small otherwise (case $1, 2, 4$). The region R for estimating ego-motion in section 2.1 is chosen as those pixels where C exceeds a pre-defined threshold.

3. EXPERIMENTS

3.1. Synthetic Example

We conducted experiments on the Yosemite sequence and a semi synthetic 3D model (with real textures) of an urban environment. Only the results on Yosemite data are presented here due to space constraints. Figure 1 shows the key image, the true depth map for the key image and the initial reference depth map (the cloud regions are not included in the experiment). The reference depth map was obtained by first smoothing the true depth map with a constant filter of size 25×25 pixels to get a coarse depth map. Gaussian noise ($\sigma = 0.07$) was then added to it. A rectangular region in the center (Figure 1(c)) of the coarse and noisy depth map was modified to a constant depth value which is equivalent to having no depth information in that region. Thus our initial depth map is coarse, noisy and lacks information at certain regions.

We use only one local iteration for depth refinement and a total of 10 global iterations. The true FOE and rotational parameters are $(0, 0.17)$ and $(0, -0.0017, 0.0003)$ respectively. Figure 2(a) shows the convergence of FOE estimates (x_f, y_f) with global iterations for a constant parallax model (CPM) and DBPM. Estimated rotational parameters using DBPM at the end of global iterations are $(0, -0.0018, 0.0005)$ which are close to the true values. The FOE estimate converges to the true value for DBPM but not for CPM indicating that our model is more robust. Figure 2(c) shows the mean confidence over the entire image using DBPM which increases as depths get refined and becomes stable. The confidence threshold for choosing R for ego-motion estimation was set to 0.3. The estimated depth maps at the end of global iterations using DBPM and CPM are shown in Figure 1(d) and (e) respectively. Qualitatively, depth map estimated using CPM is much more noisy than the one estimated using DBPM indicating that DBPM can handle noise in available data much better. Also, the artificial depth discontinuities in the center of reference depth map are not removed by CPM (the true depth map does not have those) but are handled properly by DBPM. Thus a more realistic 3D model can be obtained using DBPM when depth information is missing from certain regions. We define the relative mean square error (RMSE) between the true depth map Z_{true} and any other depth map Z as

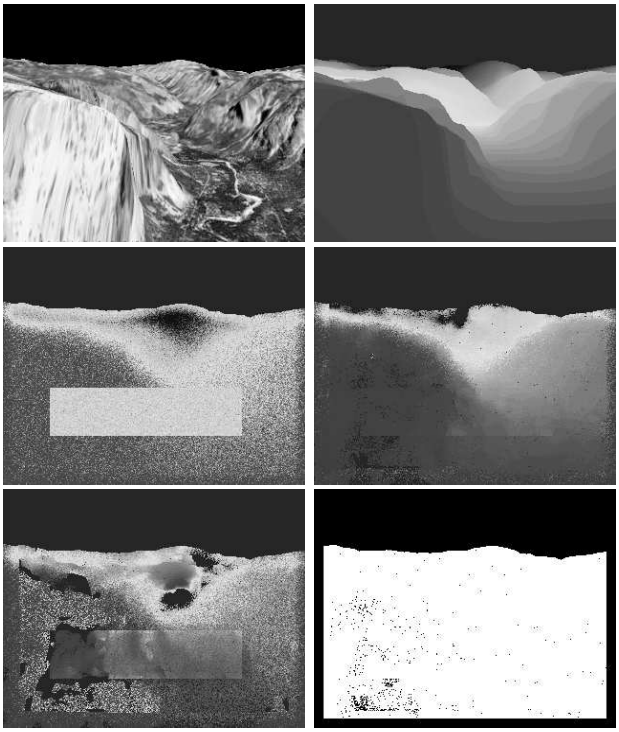

Fig. 1. (a) Key image (b) True depth map (c) Reference depth map (d) Estimated depth map using DBPM (e) Estimated depth map using CPM (f) Regions (in white) where $C \geq 0.1$ for (d)

$RMSE = \frac{100}{N} \sum_1^N (\frac{Z_{true} - Z}{Z_{true}})^2$ where N denotes the number of pixels. Table 1 gives the RMSE between the true depth map and the estimated depth maps using DBPM and CPM and shows that DBPM performs much better. These numbers are calculated at pixels where the confidence C at the end of global iterations is greater than 0.1 (shown in Figure 1(f)).

3.2. Real Example

A DEM model of (inner harbor area) Baltimore downtown was rendered in *OpenGL* and the reference depth map was obtained using the Z buffer as shown in Figure 3(b). Figure 3(a) shows the key frame from the video sequence which was captured using a Sony camcorder placed on a cart (not mounted) moving across a street. The dominant translational motion was in the camera's Z direction with vertical motion close to zero. The estimated ego-motion using DBPM and CPM are shown in Figure 2(b). Figure 2(c) shows the mean confidence over the entire image using DBPM which increases and converges with global iterations. Figure 3(c) and 3(d) shows the estimated depth maps (brighter regions are farther) using DBPM and CPM respectively. Note the correctly estimated pole in the center and the lamp post in the top right corner. The depth map estimated using DBPM is more accurate, less noisy and depth boundaries are preserved better.

4. CONCLUSIONS

An iterative algorithm is presented for estimating ego-motion and depth recovery from a noisy, coarse and partial depth map and im-

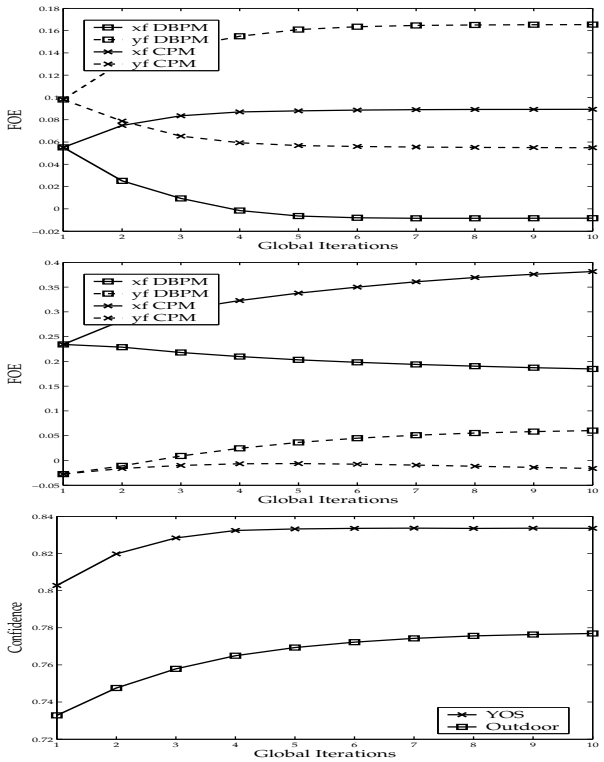

Fig. 2. Convergence of FOE estimates (a) Yosemite example (b) Real example (c) Mean confidence over the entire image for Yosemite and Real example using DBPM

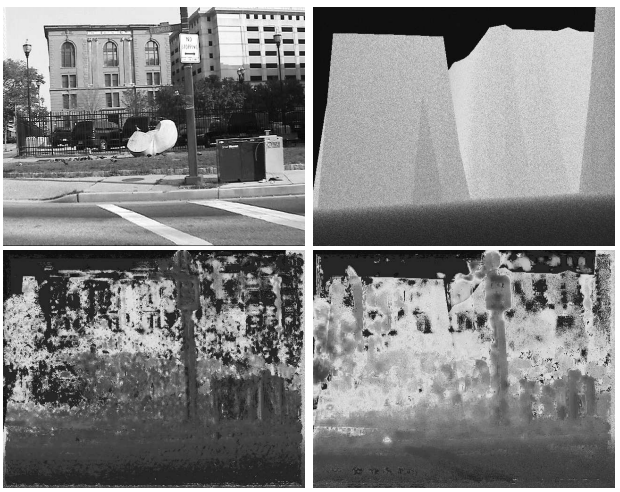

Fig. 3. Real Example (a) Key image (b) Reference depth map (c) Estimated depth map using DBPM (d) Estimated depth map using CPM

	RMSE with True Depth Map
Reference	59.40
Estimated using CPM	32.17
Estimated using DBPM	02.27

Table 1. RMSE between true depth map and the reference and estimated depth maps using DBPM and CPM

age derivatives. A new depth based parallax model is proposed for modeling the parallax field and a TLS solution along with confidence measures are derived for the model. Results and comparisons with locally smooth depth model on synthetic and real example shows the effectiveness of our approach. Future efforts will focus on extending the algorithm to multiple frames beyond the current two-frame approach.

5. REFERENCES

[1] T.S. Huang and A.N. Netravali, "Motion and structure from feature correspondences: A review," *Proceedings of the IEEE*, vol. 82, pp. 252–268, 1994.

[2] G.S. Young and R. Chellappa, "3-D motion estimation using a sequence of noisy stereo images: Models, estimation, and uniqueness results," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 735–759, 1990.

[3] H. Liu, R. Chellappa, and A. Rosenfeld, "A hierarchical approach for obtaining structure from two-frame optical flow," *Proceedings of Workshop on Motion and Video Computing*, pp. 214–219, 2002.

[4] G. Adiv, "Determining 3-d motion and structure from optical flow generated by several moving objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 7, no. 4, pp. 384–401, July 1985.

[5] Y. Aloimonos and C.M. Brown, "Direct processing of curvilinear sensor motion from a sequence of perspective images," in *CVWS84*, 1984, pp. 72–77.

[6] B.K.P. Horn and E. Weldon, "Direct methods for recovering motion," *IJCV*, pp. 51–76, 1988.

[7] K.J. Hanna, "Direct multi-resolution estimation of ego-motion and structure from motion," in *MOTION91*, 1991, pp. 156–162.

[8] S. Negahdaripour, N. Kolagani, and B.Y. Hayashi, "Direct motion stereo for passive navigation," in *CVPR92*, 1992, pp. 425–431.

[9] R. Kumar, P. Anandan, and K.J. Hanna, "Direct recovery of shape from multiple views: a parallax based approach," *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, vol. 1, pp. 685–688, 1994.

[10] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," *Proceedings of Eurpoean Conference on Computer Vision*, pp. 237–252, 1992.

[11] M. Irani, P. Anandan, and M. Cohen, "Direct recovery of planar-parallax from multiple frames," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 11, pp. 1528–1534, 2002.

[12] A.K. Agrawal and R. Chellappa, "3d model refinement using surface-parallax," *to be published in ICASSP*, 2004.

[13] H. Liu, R. Chellappa, and A. Rosenfeld, "Accurate dense optical flow estimation and segementation using adaptive structure tensors and a parametric model," *IEEE Transactions on Image Processing*, vol. 12, no. 10, pp. 1170–1180, 2003.

[14] H. Liu, T.H. Hong, M. Herman, and R. Chellappa, "A general motion model and spatio-temporal filters for computing optical flow," *IJCV*, vol. 22, pp. 141–172, 1997.