Robust Ego-Motion Estimation and 3D Model Refinement Using Surface Parallax

Amit Agrawal*, *Student Member, IEEE,* and Rama Chellappa, *Fellow, IEEE*

***Abstract*— We present an iterative algorithm for robustly estimating the ego-motion and refining and updating a coarse depth map using parametric surface parallax models and brightness derivatives extracted from an image pair. Given a coarse depth map acquired by a range-finder or extracted from a Digital Elevation Map (DEM), ego-motion is estimated by combining a global ego-motion constraint and a local brightness constancy constraint. Using the estimated camera motion and the available depth 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. The parallax magnitude is estimated using a constant parallax model (CPM) which assumes a smooth parallax field and a depth based parallax model (DBPM), which models the parallax magnitude using the given depth map. We obtain confidence measures for determining the accuracy of the estimated depth values which are used to remove regions with potentially incorrect depth estimates for robustly estimating ego-motion in subsequent iterations. Experimental results using both synthetic and real data (both indoor and outdoor sequences) illustrate the effectiveness of the proposed algorithm.**

***Index Terms*— Direct methods, 3D modeling, surface parallax**

I. Introduction

3D scene reconstruction and ego-motion estimation has been an active area of research over the past few decades. 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 image sequences to develop robust algorithms for building enhanced 3D models. The available depth information, however, is often noisy, coarse and sparse (may lack data in certain regions). In this paper, we address the problem of using such low quality sparse depth information along with intensity images to estimate the ego-motion and the depth map of the scene.

Majority of work on ego-motion estimation assume that correspondences between image features or tokens are given and focus on recovering structure and motion from these features [1] [2] [3] [4] [5] [6] [7] [8] [9]. Another class of methods [10] [11] [12] [13] assume that optical flow or a set of dense correspondences is available between image frames and recover dense 3D structure using them. However, recovering robust optical flow itself is a non-trivial problem. Direct Methods [14] [15] [16] [17] [18] [19] [20] [21] try to combine the correspondence estimation problem with structure recovery by simultaneously estimating both structure and motion and avoiding the intermediate step of computing flow or correspondences between features. These techniques minimize the deviation of brightness change model (constant brightness model or generalized dynamic image model [22]) with respect to structure and motion parameters.

The algorithm presented in this paper comes under the category of direct methods. Several direct methods focus on expressing the image motion of rigid objects as a sum of translational and rotational fields. Techniques such as [16] [21] recover the 3D structure relative to the camera. Alternative approaches such as "plane + parallax" [23] [18] [24] [25] [26] [27] recover the 3D structure relative to a reference plane. Most of the "plane+parallax" approaches assume the presence of a dominant plane in the scene or a piece-wise planar model [28] [29]. In these methods, the homography for the dominant planar surface is estimated. This homography encapsulates the rotational motion of the camera and the calibration effects. The residual image motion is then an epipolar field and is due to deviations of the scene structure from the planar surface. The scene structure can thus be refined by estimating the residual motion or the parallax. However, the above assumption is not valid in several scenarios. In this paper, we show how any non-planar surface can be used to recover dense 3D structure, thereby not requiring the assumption that a piecewise planar model or a dominant planar surface be present in the scene. The approach presented here can work with general 3D scenes.

In addition, most of the previous approaches assume locally smooth depth models for estimating depths [16] [19] or small depth variations compared to the distance from the camera [10] [17]. However, these assumptions are violated when the depth variations are 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 modeling the parallax field on depths can deal with such cases. Besides, 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 by incorporating a suitable confidence measure in estimating depths.

Recently, there has been considerable research on sensitivity and robustness of existing algorithms in computer vision for SfM, optical flow etc. Kearney et. al. [30] did an error analysis of gradient based methods for optical flow. Weng et.

emails: [aagrawal,rama]@cfar.umd.edu. The authors are with Center for Automation Research, University of Maryland, College Park, MD 20742. 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.

al. [5] use first-order perturbations in the inputs to estimate the standard deviation of the error in the reconstructed scene from perspective views. Broida and Chellappa [31] derived Cramer-Rao lower bounds on the estimated error variance of the structure and motion parameters from a sequence of images under perspective projection. Young and Chellappa [32] derived bounds on the estimation error for structure and motion parameters from two images using optical flow. Oliensis [33] did a least squares error analysis for SfM as a function of camera motion. Soatto [34] proposed a global error analysis by casting SfM as the minimization of a high-dimensional quadratic cost function. In section III, we identify local sources of error (in the depth refinement phase using the surface parallax approach) such as errors due to sampling and discretization in image gradients, errors near the Focus of Expansion (FOE), and those due to homogeneous regions and regions where local edge structure is aligned towards the parallax direction. Errors in camera motion estimation and noise in the available depth map will also create problems. Fortunately, most of the above cases can be identified in the image. The use of a eigen-value techniques allows us to define a reasonable confidence measure in terms of eigen-values, thus providing a measure of the quality of the depth estimates.

The rest of the paper is organized as follows. Section II describes the algorithm. Section III identifies local sources of error and suggests ways to handle them. Section IV presents experimental results using both synthetic and real 3D models (both indoor and outdoor image sequences). Qualitative and quantitative comparisons of the estimated depth map and ego-motion using CPM and DBPM and with previous algorithm [16] are presented.

II. Algorithm

The proposed method is a direct approach that uses two intensity images (referred to as *key* and *offset* frames) and an initial coarse, noisy and incomplete 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.

Let $\mathbf{p} = (x, y)$ denote an image pixel and t denote the time index. Assuming brightness constancy, we have

$$I(\mathbf{p}, t) = I(\mathbf{p} - \mathbf{u}, t - 1) \tag{1}$$

where $I(\mathbf{p}, t)$ and $I(\mathbf{p}, t-1)$ denote the key and offset frames respectively. The 2D image motion $\mathbf{u}$ is given by [15]

$$\mathbf{u} = A\widetilde{Z}T + B\Omega \tag{2}$$

where $B = \begin{bmatrix} \frac{xy}{f} & -(f + \frac{x^2}{f}) & y \\ (f + \frac{y^2}{f}) & -\frac{xy}{f} & -x \end{bmatrix}$, $A = \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix}$, $\widetilde{Z} = \frac{1}{Z}$ denotes the inverse depths, f denotes the focal length (which we assume is known) and (T,Ω) denote the translational and rotational velocities of the camera. For estimating ego-motion and depth, we minimize the deviations from the brightness constancy equation

$$E = \sum_R (I(\mathbf{p}, t) - I(\mathbf{p} - \mathbf{u}, t - 1))^2 \tag{3}$$

over suitable image regions R.

A way to minimize (3) is to perform iterative Gauss-Newton minimization (we call these iterations *local iterations*) which uses a first order expansion of individual quantities before squaring in error term E. Let i denote the global iteration index, $\mathbf{u}_i$ denote the current estimate of the flow field during the i^{th} global iteration (obtained from current depth and motion estimates using (2)). Let $\mathbf{du}_i^m$ and $\mathbf{du}_i^Z$ denote the incremental 2D motion for a local iteration due to motion refinement and depth refinement respectively. The appropriate motion (or depth) refinement can be estimated [19] by minimizing

$$E(\mathbf{du}_i^{m \text{ or } Z}) = \sum_R (\nabla I^T \mathbf{du}_i^{m \text{ or } Z} + \Delta I)^2 \tag{4}$$

with respect to $\mathbf{du}_i^{m \text{ or } Z}$, where $\nabla I = [I_x, I_y]^T$ denotes the spatial image derivatives and $\Delta I = I(\mathbf{p}, t) - I(\mathbf{p} - \mathbf{u}_i, t - 1)$ denotes the difference of the key image and the warped offset image according to $\mathbf{u}_i$ which is obtained from current depth and motion estimates.

A. Ego-Motion Estimation Given a Depth Map

Let Z_i denote the current estimate of the depth map from the previous global iteration (for the first global iteration we use the reference depth map). To estimate the ego-motion, we minimize (3) 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 II-E (for the first global iteration we use the entire image region).

Let $m_i = [T_i, \Omega_i]^T$ 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 $dT, d\Omega$ be the incremental ego-motion update for a local iteration. Using (2), we have

$$\mathbf{du}_i^m = A\widetilde{Z}_i dT + Bd\Omega = \begin{bmatrix} A\widetilde{Z}_i & B \end{bmatrix} \times dm \tag{5}$$

where $dm = [dT, d\Omega]^T$ denotes the incremental ego-motion. Substituting the above equation in (4), we get

$$E(\mathbf{du}_i^m) = \sum_R (\nabla I^T \begin{bmatrix} A\widetilde{Z}_i & B \end{bmatrix} dm + \Delta I)^2 \tag{6}$$

where ΔI is calculated using $\mathbf{u}_i$ obtained from m_i and Z_i. This is a linear system in dm and a least square solution is obtained. The update dm is added to the current motion estimate m_i to get a refined estimate. Thus at each local iteration, m_i is refined, a new value of ΔI is obtained using refined m_i and Z_i and (6) is minimized to obtain further refinement dm. The quality of fit can be determined by evaluating (3) using refined m_i and Z_i. The local iterations are performed until the error E in (3) stops decreasing or the change in motion dm falls below a pre-define threshold (10^{-6}).

B. Depth Refinement using Ego-motion

Let T_i, Ω_i denote the current ego-motion estimate and Z_i denote the available depth map estimate. Let dZ_i be the incremental depth map estimate for the i^{th} global iteration and $Z_{i+1} = Z_i + dZ_i$ be the refined depth map. Using (2), the incremental 2D motion can be written as

$$\mathbf{du}_i^Z = A(\widetilde{Z}_{i+1} - \widetilde{Z}_i)T_i \tag{7}$$

where $\widetilde{Z}_{i+1} = \frac{1}{Z_{i+1}}, \widetilde{Z}_i = \frac{1}{Z_i}$. 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 camera motion T_i from previous ego-motion estimate, we can constrain the direction of the parallax field.

First, let $T_z \neq 0$. Defining the FOE as $(x_f = f\frac{T_x}{T_z}, y_f = f\frac{T_y}{T_z})$, we constrain the direction of the parallax field to lie along the epipolar direction. Thus for each pixel (x, y) we write

$$\mathbf{du}_i^Z(x, y) = \beta \begin{bmatrix} x - x_f \\ y - y_f \end{bmatrix} \tag{8}$$

where $[x - x_f, y - y_f]^T$ denotes the parallax direction and β denotes the parallax magnitude. Expanding AT_i in (7), we get

$$\mathbf{du}_i^Z(x, y) = T_z \begin{bmatrix} x - x_f \\ y - y_f \end{bmatrix} (\widetilde{Z}_{i+1} - \widetilde{Z}_i)$$

Comparing the above equation with (8), one obtains

$$\beta = T_z(\widetilde{Z}_{i+1} - \widetilde{Z}_i) \tag{9}$$

Now consider the case $T_z = 0$. The epipolar field in that case is oriented along the 2D direction $[T_x, T_y]^T$. For example, if the camera is moving along the X axis, the epipolar field will be horizontal. Here we write $\mathbf{du}_i^Z(x, y)$ as

$$\mathbf{du}_i^Z(x, y) = \beta \begin{bmatrix} T_x \\ T_y \end{bmatrix} \tag{10}$$

Expanding AT_i in (7) using $T_z = 0$, we get

$$\mathbf{du}_i^Z(x, y) = -f \begin{bmatrix} T_x \\ T_y \end{bmatrix} (\widetilde{Z}_{i+1} - \widetilde{Z}_i)$$

Comparing the above equation with (10), one obtains

$$\beta = -f(\widetilde{Z}_{i+1} - \widetilde{Z}_i) \tag{11}$$

We first estimate β as follows. Using (8) (or (10) for $T_z = 0$), (4) can be written as

$$E = \sum_R (I_d\beta + \Delta I)^2 \tag{12}$$

where I_d denotes the projection of the intensity gradient along the parallax direction. The region R for depth refinement is chosen to be a local neighborhood of $N \times N$ pixels. We first minimize (12) to get an estimate of β and then use β to obtain refined depths $\widetilde{Z}_{i+1}$ using (11) or (9), depending on whether T_z is zero or not. Next, we describe how to obtain the parallax magnitude using various parallax models.

C. Estimating Parallax Magnitude

We estimate the parallax magnitude pixel by pixel. For each pixel, (12) can be minimized with respect to β giving a least squares (LS) solution. LS solution for solving a linear system $Ax = b$ to estimate x assumes noise to be present in b (ΔI here) only. However, a better estimate based on total least squares (TLS) [35] can be obtained assuming both A and b (I_d and ΔI here) to be noisy (see for e.g. optical flow computation in [36] [37] [38]).

The TLS solution can be formulated as minimizing

$$J = \left\langle [g^T\gamma]^2 \right\rangle \tag{13}$$

with respect to γ where $g = [I_d, \Delta I]^T$, $\gamma = [\beta_1, \beta_2]^T$ and $<>$ defines the mean operator

$$< f(\overline{x}, \overline{y}) >= \int_{-\infty}^{\infty} w(x - \overline{x}, y - \overline{y}) f(x, y) dx dy \tag{14}$$

where w is a windowing function. The parallax magnitude β is then given by $\beta = \frac{\beta_1}{\beta_2}$.

1) Solving for CPM: We assume β to be constant over the region R leading to the *constant parallax model*. This is similar in spirit to having a smoothness constraint on depths by assuming a smooth depth model (constant or planar) (as in [16] [19]) or assuming constant dZ_i over the neighborhood to estimate Z_{i+1}. 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

$$J = \left\langle \gamma^T g g^T \gamma \right\rangle + \lambda(1 - \gamma^T\gamma) = \gamma^T G\gamma + \lambda(1 - \gamma^T\gamma)$$
$$G = \left\langle gg^T \right\rangle = \begin{bmatrix} < I_d^2 > & < I_d\Delta I > \\ < I_d\Delta I > & < \Delta I^2 > \end{bmatrix} \tag{15}$$

Differentiating with respect to γ, we get $G\gamma = \lambda\gamma$. Since G is a 2×2 real symmetric matrix, there will be two valid eigen-value/eigen-vector pairs. Let $\lambda_1 \geq \lambda_2$ be the valid eigen-values. The eigen-vector corresponding to λ_2 will be the solution for γ.

2) Solving for DBPM: The assumption of a locally smooth depth model is violated at depth boundaries when significant depth variations are present. Also, the effect of noise in the available depth map estimate (from a range finder or DEM) may require a non-smooth depth refinement within the neighborhood. Thus, in such cases, the parallax magnitude is not smooth over the neighborhood. From (9) and (11), we observe that the parallax magnitude β depends on $\widetilde{Z}_{i+1} - \widetilde{Z}_i = \frac{1}{Z_{i+1}} - \frac{1}{Z_i} = \frac{-dZ_i}{Z_i^2}$. Noting that the parallax magnitude depends on inverse depths, DBPM is defined as

$$\beta = a_0 + \frac{a_1}{Z_i} + \frac{a_2}{Z_i^2} \tag{16}$$

where the parameters a_0, a_1 and a_2 are assumed to be constant within the neighborhood. Note that even though a parametric model is used to model β, 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.

Define $B = \begin{bmatrix} 1 & \widetilde{Z}_i & \widetilde{Z}_i^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \widetilde{Z}_i & \widetilde{Z}_i^2 \end{bmatrix}$. Using DBPM, we write

$$\gamma = Bp \tag{17}$$

where $p = [a_0, a_1, \ldots, a_5]^T$ denotes the model parameters to be estimated. 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

$$J = \langle \gamma^T g g^T \gamma \rangle + \lambda(1 - \gamma^T\gamma) = p^T T p + \lambda(1 - p^T D P) \tag{18}$$

where $T = \langle B^T g g^T B \rangle$ and $D = B^T B$. Differentiating with respect to p, we get $Tp = \lambda Dp$ subject to $p^T Dp = 1$. Thus the minimization problem is converted into the solving a generalized eigenvalue system. 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. Using the estimate p and (17), one can obtain γ and hence β.

D. Regions with No Prior Depth Information

The initial reference depth obtained from a range sensor or DEM may be lacking in information in certain regions. For e.g. in some regions no prior depth information may be available. A reasonable assumption is to assume a constant depth value for such regions in the reference depth map. From (16), we observe that if the available depths are constant over the region R, the DBPM reduces to CPM. So for such regions, one can selectively use the CPM in our formulation. We will show in section IV that for scenes with sufficiently low depth variations, no prior depth information may be required by the algorithm and good results can be obtained starting from a flat initial depth.

E. Confidence Measures

Confidence measures based on eigen-values and/or condition number have been proposed in [37] [39]. We use $C = (\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2})^2$ as the confidence measure for depth estimation. When the solution is estimated reliably, λ_2 will be close to zero and λ_1 will be sufficiently greater than zero and hence confidence will be close to one. In addition, we assign a confidence of zero whenever the magnitude of I_d falls below some pre-defined threshold to avoid ill-conditioning of the system. Thus homogeneous regions, regions where local edge structure is aligned along the parallax direction and regions near FOE are assigned a confidence of zero. Note that the above confidence measure is normalized between 0 and 1. When both λ_1 and λ_2 are close to zero, it may be unreliable. Hence, a threshold on the sum of eigen-values is used to identify such cases and confidence measure at those pixels is set to zero. The region R for estimating ego-motion in section II-A is composed of those pixels where C exceeds a pre-defined threshold.

F. Algorithm Outline

1) Get the initial reference depth map Z_0, key and offset frames. Set the global iteration index $i = 1$.
2) Estimate the camera motion m_i using Z_{i-1} (as explained in II-A)
3) Refine the depths using m_i and Z_0 using DBPM or CPM. Let the refined depths be Z_i. Obtain confidence measures for depth estimates. Set $i \rightarrow i + 1$.
4) Repeat step 2 by setting R to those regions in image where the confidence in depth estimates is greater than a pre-defined threshold. Repeat step 3.
5) Stop when the maximum iterations are reached or ego-motion parameters converge.

III. Error Identification

We now identify various sources of error that influence the depth estimates, including camera motions for which algorithm can fail. Errors consists of those due to discretization of the derivative operator, sampling measurement errors and statistical errors caused by noise in the imaging process and in the reference depth map (due to range sensor noise for example.)

A. Gradient Measurement Error

Errors in spatial gradient measurement are related to higher order spatial image gradients. First order discrete difference operators introduce a large error in spatial gradients I_x and I_y. Error in I_x, $\epsilon_{I_x} \approx (\Delta x) I_{xx}$ for a forward differencing operator [30] and $\epsilon_{I_x} \approx (\Delta x)^2 I_{xxx}$ [38] for the central differecing operator. Better derivative operators include series designed operators [40] with a large support for accurate derivative computation. Since gradient errors are related to higher order derivatives, this emphasizes the need for proper prior smoothing of images. In addition, the errors in gradients will also be corrupted by the errors in brightness estimates and sampling errors. In [41], it was shown that the total least square solution for computing 1-D optical flow is unbiased as compared to a least square solution if the noise is isotropic in all gradients which is another advantage of having a total least squares solution.

ΔI in (4) can be regarded as a directional derivative in the direction $\mathbf{u}$ (if $\mathbf{u} = 0$, it reduces to a temporal derivative). The error in estimating the directional derivative grows as the square of the flow magnitude [30] (parallax magnitude in our case). Thus we expect regions with high parallax magnitude to have high error in estimated depths.

B. Non-uniform Parallax and Ill-conditioning

The assumption of constant parallax magnitude within a small neighborhood is violated at depth boundaries and with errors in available depth estimates. As a result, local optimization will provide inferior results. This can be alleviated using DBPM where the parallax magnitude is derived from the depth values itself.

The accuracy of the estimated parallax magnitude depends on errors in I_d, ΔI and error propagation of the linear system. The errors in I_d depend on spatial image derivatives and camera translational motion. When $T_z \neq 0$, $I_d = I_x(x - x_f) + I_y(y - y_f)$. The system will be ill-conditioned when I_d is close to zero. In such cases, a small error in these values can cause a large error in β. Consider the following scenarios

1) Homogeneous regions: No intensity variation in spatial direction. $I_x = I_y = 0$.
2) Spatial intensity gradient is in a direction perpendicular to the parallax direction, i.e. $[I_x, I_y]^T \perp [x - x_f, y - y_f]^T$. Physically, this corresponds to the case when local edge structure is aligned along the direction of FOE.
3) Regions near FOE: $x \approx x_f, y \approx y_f$.

For all the above cases, I_d is close to zero and hence no reliable solution can be obtained. These conditions, however can be identified in the image. If $T_z = 0$, $I_d = I_x T_x + I_y T_y$, and hence, the first two cases mentioned above will lead to errors.

The conditioning can be improved by using a large neighborhood size. The risk involved in using a large neighborhood is that the parallax model may not be valid. Thus the neighborhood size N defining the window function w should be carefully chosen. We use the adaptive windowing approach presented in [37]. Initially, a small window size is used to estimate the parallax and is adaptively increased until the confidence measure (as described in section II-E) stops increasing.

C. Effect of Camera Motion

Since the proposed algorithm is a parallax based algorithm, camera motion must include translation. Thus the case of pure rotation falls into a degenerate case. Hence, if the camera translation is close to zero, the estimated depths will not reliable. In all cases, errors in camera motion estimate will lead to errors in parallax estimation. Let e_x denote the error in any quantity x. Then the error in I_d due to errors in camera motion is

$$e_{I_d} = -I_x e_{x_f} - I_y e_{y_f} \tag{19}$$

The errors in ΔI will occur through the errors in optical flow estimates $\mathbf{u}$. Let $e_{\mathbf{u}}$ be the errors in optical flow estimates due to errors in camera motion. The estimated directional derivative will be

$$\widehat{\Delta I} = I(\mathbf{p}, t) - I(\mathbf{p} - \mathbf{u} - e_{\mathbf{u}}, t - 1) \tag{20}$$

Assuming $e_{\mathbf{u}}$ to be small, we can approximate

$$\widehat{\Delta I} = I(\mathbf{p} + e_{\mathbf{u}}, t) - I(\mathbf{p} - \mathbf{u}_i, t - 1) = \Delta I + \nabla I e_{\mathbf{u}}$$

Thus the error in ΔI is $e_{\Delta I} = \widehat{\Delta I} - \Delta I = \nabla I e_{\mathbf{u}}$. Thus, errors in the camera motion estimate introduce an error in both I_d and ΔI, and these errors depend on spatial image gradients. Thus, we see that there are several factors which interact in a complex way to determine the accuracy of the estimated parallax magnitude and hence depths.

IV. Experiments

We conducted experiments using both synthetic and real images. For synthetic images, we present results on the Yosemite sequence (referred to as *YOS*) and a 3D model of an urban environment (referred to as *3DS*). For real images, results on an outdoor and indoor sequence are presented (referred to as *Outdoor* and *LABS* respectively). The experiments show the effectiveness of the algorithm for different camera motion in terms of the FOE being in the image region (*YOS*, *Outdoor*), FOE outside the image region (*3DS*), FOE at infinity (*LABS*) and scene structure in terms of high depth variability (*YOS*, *3DS*, *Outdoor*), low depth variability (*LABS*). We have also implemented the multi-resolution algorithm described in [16] (referred to as HANNA and provide comparisons with it. For HANNA, we use a fixed 5×5 neighborhood for local optimization (as in [16]) with 3 levels of resolution. The percentage depth error is defined as

$$\frac{100}{N} \sum_{1}^{N} \left(\frac{\text{true depth} - \text{computed depth}}{\text{true depth}}\right)^2 \tag{21}$$

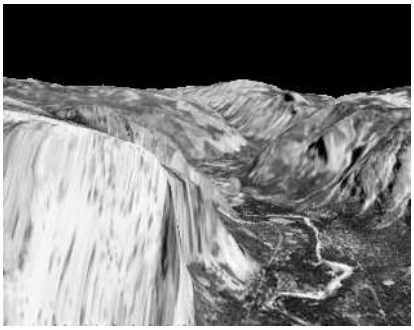
where N denotes the number of pixels following [16]. In all experiments, one local iteration was done for depth refinement. The confidence threshold for choosing the region R for ego-motion estimation was set to 0.3. In all examples, we smooth the images using a Gaussian filter with standard deviation of 1 along the spatial axes. For computing derivatives along the spatial axes, a series-designed filter of radius 5 was used whose filter coefficients are $[0.0036, -0.0381, 0.2, -0.8, 0, 0.8, -0.2, 0.0381, -0.0036]$. Thus gradient estimates and hence the depth estimates will be unreliable at the periphery of the image due to the lack of sufficient information in computing derivatives. For synthetic sequences, the error depth maps are also shown (gray color coded with white indicating large errors and black indicating zero error).

A. Yosemite

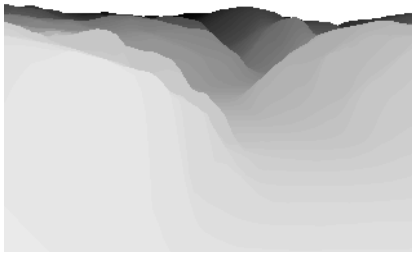
We present result on Yosemite sequence. The cloud regions were not included in the experiment. Figs. 1(a), 1(b) and 1(c) show the key image, the true depth map for the key image and the initial reference depth map respectively. The depth map is color coded (darker regions are farther from the camera). As a coarse depth map from an independent source was not available, we obtain the initial reference depth map from the true depth map as follows.

The initial reference depth map was obtained by first smoothing the true depth map with a constant filter of size 25×25 pixels to get a highly coarse depth map. Gaussian noise was then added to it. A rectangular region in the center (Fig. 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. In addition, the rectangular region in the center also introduces significant artificial depth discontinuities (at the boundaries of the rectangle) in the depth map which are not present in the true depth map. Thus the initial reference depth map is coarse, noisy, lacks information in certain regions and has depth discontinuities.

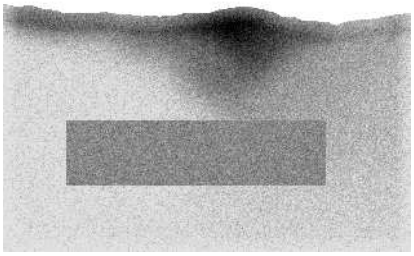
We performed a total of 10 global iterations. The true FOE and rotational velocities (in radians) are $(0, 0.17)$ and $(0, -0.0017, 0.0003)$ respectively. Figs. 2(a) and 2(b) shows the convergence of FOE estimates (x_f, y_f) with global iterations using CPM, DBPM and HANNA respectively. The FOE estimate converges to the true value for DBPM. Estimated rotational parameters using DBPM at the end of global iterations are $(0, -0.0018, 0.0005)$ which are close to the true



(a) Key image

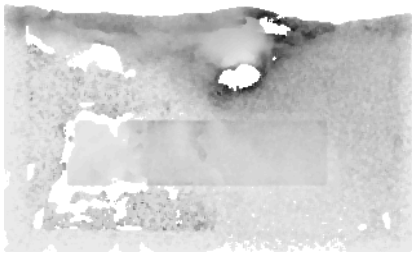


(b) True depth map

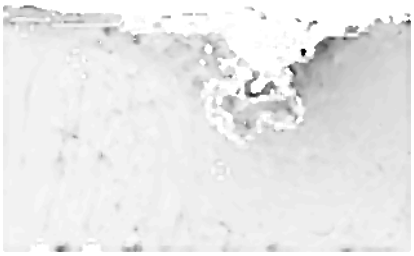


(c) Reference depth map

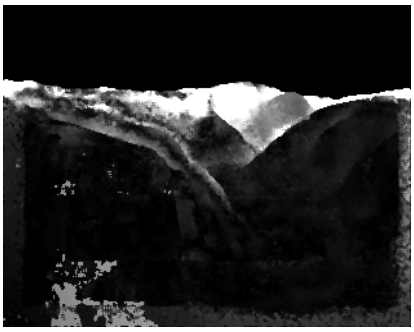
(d) Estimated depth map using DBPM



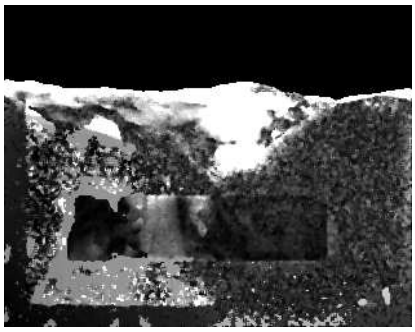
(e) Estimated depth map using CPM



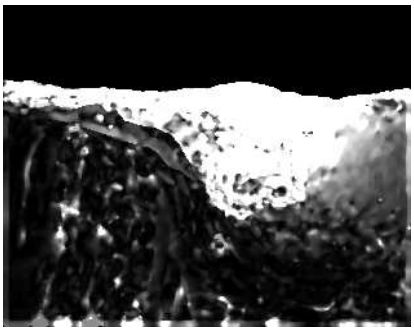
(f) Estimated depth map using HANNA



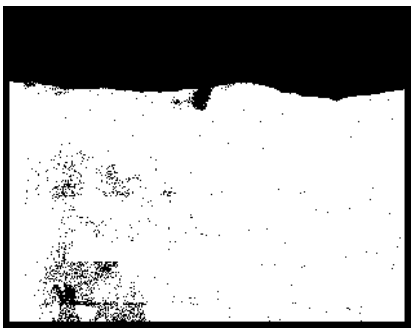
(g) Error depth map for DBPM



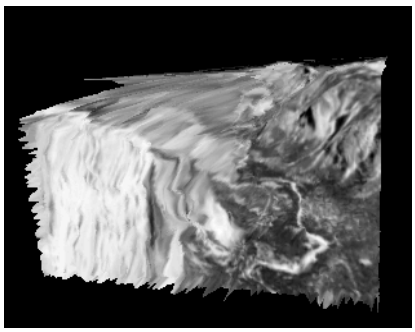
(h) Error depth map for CPM



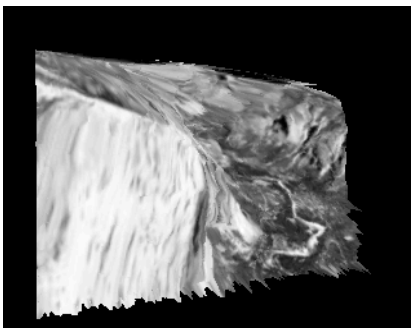
(i) Error depth map for HANNA



(j) Regions (in white) where $C \geq 0.1$ for (d)

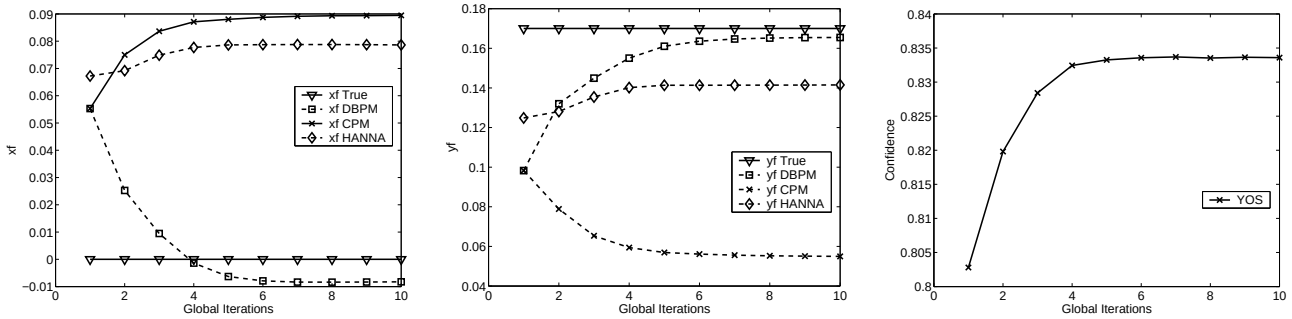


(k) Novel View



(l) Novel View

Fig. 1. YOS: (k and l) Rendered scene from novel viewpoints using depth map estimated using DBPM.



(a) (b) (c)

Fig. 2. YOS: (a and b) Convergence of x_f and y_f estimates. (c) Mean confidence over the entire image using DBPM.

TABLE I

YOS: PERCENTAGE DEPTH ERRORS FOR INITIAL COARSE DEPTH MAP AND ESTIMATED DEPTH MAPS USING DBPM, CPM AND HANNA.

	Percentage depth error
Initial coarse depth map	59.40
Estimated using CPM	32.17
Estimated using HANNA	26.01
Estimated using DBPM	02.27

TABLE II

3DS: PERCENTAGE DEPTH ERROR FOR THE INITIAL COARSE DEPTH MAP AND ESTIMATED DEPTH MAPS USING DBPM, CPM AND HANNA.

	Percentage depth error
Initial coarse depth map	47.09
Estimated using HANNA	43.49
Estimated using CPM	10.72
Estimated using DBPM	03.56

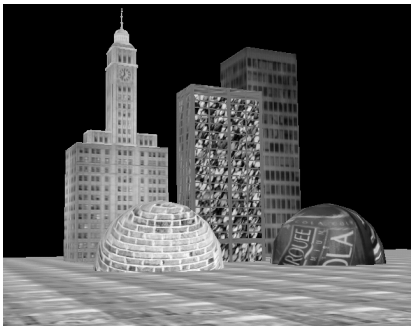
values. Fig. 2(c) shows the mean confidence over the entire image using DBPM which increases as depths get refined and become stable. Thus the proposed confidence measure is indeed a good indication of the reliability of the depth estimates. The estimated depth maps at the end of global iterations using DBPM, CPM and HANNA are shown in Figs. 1(d), 1(e) and 1(f) respectively. Figs. 1(g), 1(h) and 1(i) shows the error depth maps for DBPM, CPM and HANNA respectively. Qualitatively, the depth map estimated using DBPM is better. Also, the artificial depth discontinuities in the center of reference depth map are not removed by CPM but are handled properly by DBPM. Table I gives the percentage depth error between the true depth map and the estimated depth maps using DBPM, CPM and HANNA. These numbers are calculated at pixels where the confidence C at the end of global iterations is greater than 0.1 (shown in Fig. 1(j)). The estimated depth map obtained using DBPM was rendered in OpenGL followed by texture mapping. Figs. 1(k) and 1(l) shows two novel views of the rendered texture mapped 3D model.

B. Urban 3D Model

A semi-synthetic 3D model (with real textures) of an urban environment was rendered in OpenGL. The synthetic 3D model consists of buildings and objects in front of the buildings. We simulate a sequence of images by moving a virtual camera in the scene. The depth maps were obtained from the OpenGL Z buffer. Figs. 3(a), 3(b) and 3(c) show the key image, the true depth map for the key image and the initial coarse and noisy reference depth map respectively. The depth map is color coded (darker regions are farther from the camera). Note that there is no information for certain image regions in reference depth map, namely, the part of the building in the center of the depth map and the two spheres in front. For all such regions a constant depth value was chosen as the initial depth estimate. We performed a total of 15 global iterations. Fig. 4(a) and 4(b) shows the convergence of FOE values for CPM, DBPM and HANNA and Fig. 4(c) shows the mean confidence over the entire image for DBPM. The true FOE and the rotational velocities for this example are $(0.20, -0.39)$ and $(0.0018, -0.0017, 0.0020)$ respectively. The final estimated FOE values and rotational parameters using DBPM are $(0.21, -0.34)$ and $(0.0016, -0.0017, 0.0021)$ respectively which are close to true values. Figs. 3(d), 3(e) and 3(f) show the estimated depth map using DBPM, CPM and HANNA respectively (regions in white indicating background or negative depths). Figs. 3(g), 3(h) and 3(i) show the error depth maps for DBPM, CPM and HANNA respectively. Table II gives the percentage depth error between the true depth map and the initial sparse depth map. The texture mapped 3D model rendered from depth map estimated using DBPM is shown in Fig. 3(k).

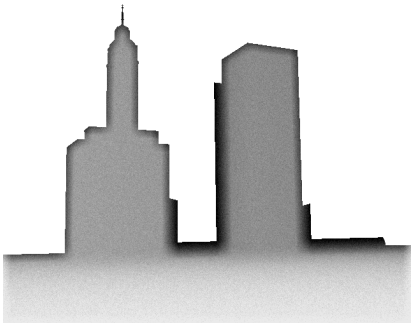
C. Outdoor Sequence

A DEM model of Baltimore downtown (inner harbor area) was rendered in *OpenGL* and the reference depth map was obtained using the Z buffer as shown in Fig. 5(b). The depth map is color coded (brighter regions are farther from the camera). The regions where no depth information is available is shown in black. Fig. 5(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. Thus the camera motion was not very smooth. The dominant translational motion was in the camera's Z direction



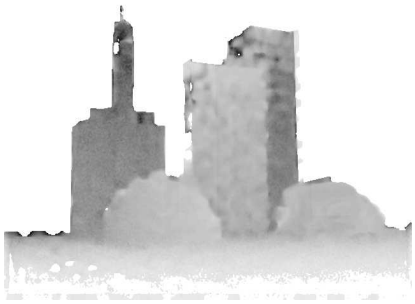
(a) Key image

(b) True depth map

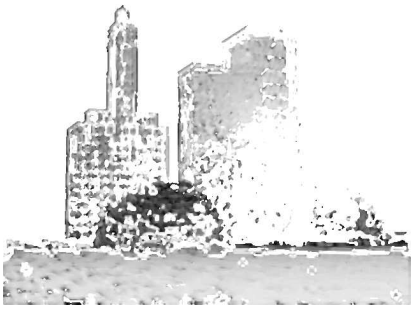


(c) Reference depth map

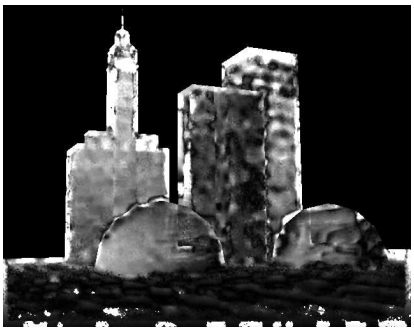
(d) Estimated depth map using DBPM



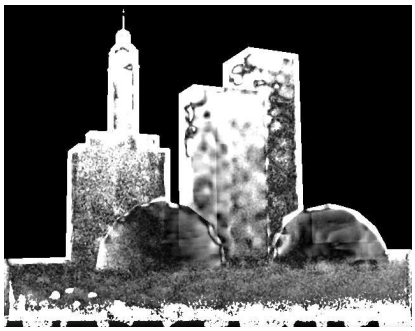
(e) Estimated depth map using CPM



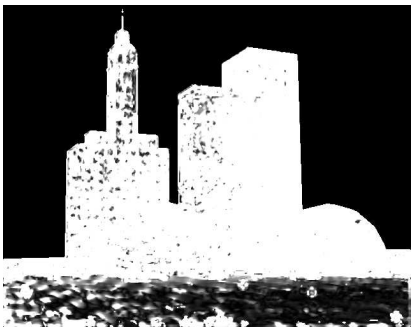
(f) Estimated depth map using HANNA



(g) Error depth map for DBPM

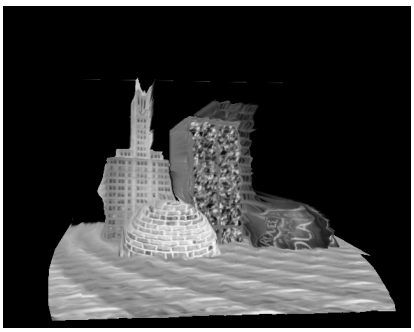


(h) Error depth map for CPM



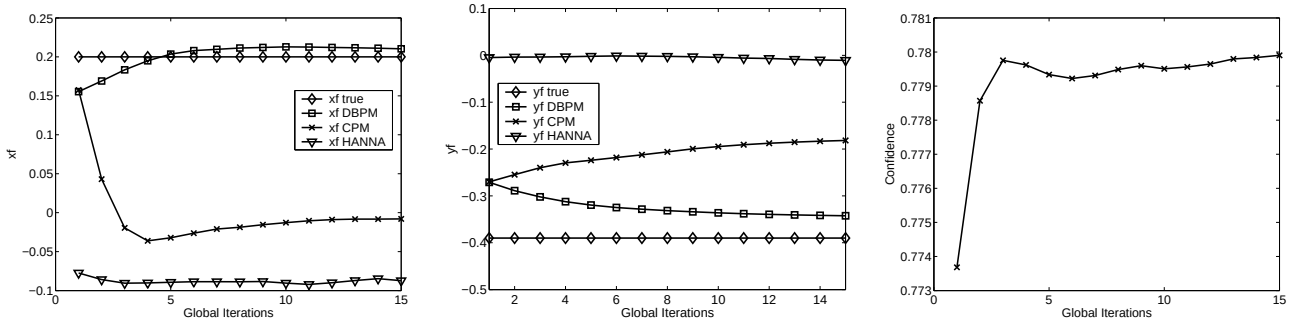
(i) Error depth map for HANNA

(j) Regions (in white) where $C \geq 0.1$ for (d)



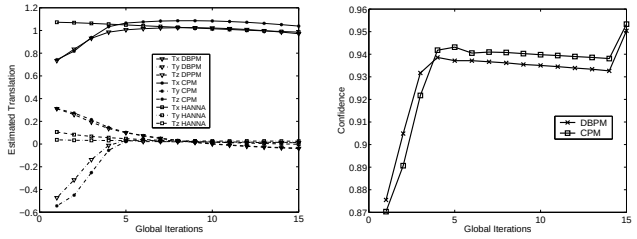
(k) Texture mapped 3D model

Fig. 3. 3DS



(a) (b) (c)

Fig. 4. 3DS: (a and b) Convergence of x_f and y_f estimates. (c) Mean confidence over the entire image using DBPM.



(a) (b)

Fig. 7. LABS: (a) Convergence of ego-motion parameters using DBPM, CPM and HANNA. (b) Mean confidence over the whole image using DBPM and CPM.

with vertical motion close to zero.

The estimated FOE parameters using DBPM, CPM and HANNA are shown in Figs. 6(a) and 6(b). Fig. 6(c) shows the mean confidence over the entire image using DBPM. Figs. 5(d), 5(e) and 5(f) show the estimated depth maps (brighter regions are farther) using DBPM, CPM and HANNA respectively. Note the correctly estimated pole and box in the center and the lamp post in the top right corner. Fig. 5(c) shows the texture mapped 3D model rendered from the depth map estimated using DBPM.

D. LABS Sequence

A video sequence of several toy objects was taken in a lab. The dominant camera motion was in the X direction. Fig. 8(d) shows the key image from the sequence. For this sequence, we did not have any prior depth information for the entire image. Also, since this is an indoor lab sequence, the variation in the scene depth is small. Therefore, the reference depth map was chosen to be a constant all over the image. A total of 15 global iterations were performed. Fig. 7(a) shows the convergence of ego-motion parameters with global iterations for CPM. The final estimated ego-motion parameters using CPM were $T_x = 1.04, T_y = 0.01, T_z = -0.03, w_x = 0, w_y = 0.0012, w_z = 0$. Thus the ego-motion parameters were estimated correctly (since T_y and T_z are zero, T_x and depths can be recovered only up to a scale factor). Figs. 8(b) shows the estimated depth map using CPM (darker regions are farther).

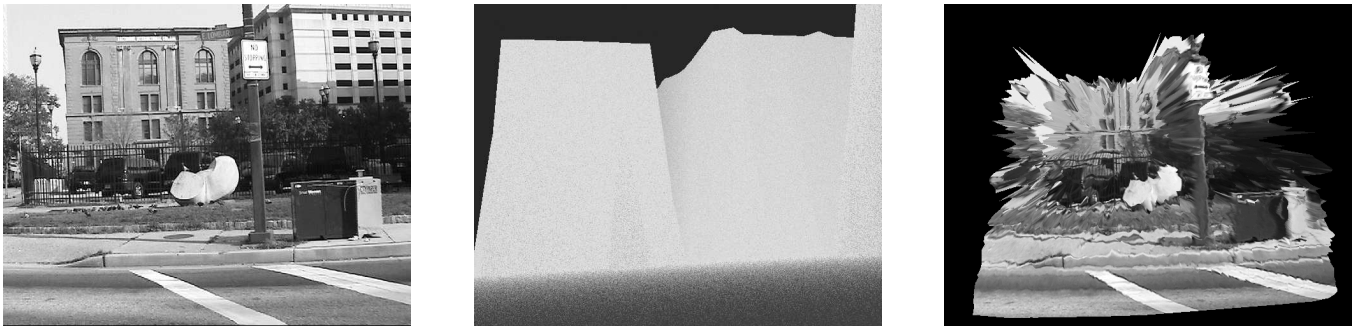
Since the initial depth map Z_0 is a constant over the entire image, as explained in section II-D, the DBPM simplifies to CPM. However, for the sake of completeness and comparison, we estimate the ego-motion and depth map using DBPM by adding a small amount of noise in the reference depth map. Fig. 7(a) shows the convergence of the ego-motion parameters with global iterations for DBPM and HANNA. Fig. 7(b) shows the mean confidence measure over the entire image using CPM and DBPM which increases as depths get refined. The final estimated ego-motion parameters using the DBPM were $T_x = 0.97, T_y = 0.01, T_z = -0.03, w_x = 0, w_y = 0.0011, w_z = 0$, which are close to the ego-motion estimates obtained using CPM. Figs. 8(a) and 8(c) shows the estimated depth map using DBPM and HANNA respectively (darker regions are farther). Note the finely extracted depth boundaries for different objects for both DBPM and CPM as compared to HANNA. Figs. 8(e) and 8(f) show two novel views of the texture mapped 3D model rendered in OpenGL from the depth map estimated using DBPM. Thus, this example shows that for indoor environments, our algorithm works well using both CPM and DBPM.

V. Conclusions

An iterative algorithm is presented for estimating ego-motion and depth recovery from a noisy, coarse and sparse depth map and image derivatives. A constant parallax model and a new depth based parallax model for handling significant depth variations and noise was described for modeling the parallax field and a total least squares solution along with confidence measures are derived for both models. Results and comparisons on synthetic and real sequences (indoor and outdoor sequences) shows the effectiveness of our approach for various camera motions and scene structure. In the presence of significant depth variations and noise in depth estimates, the depth based parallax model performs much better. When the depth variations in the scene is less, an initial flat depth can be used without the need for any prior depth information. Future efforts will focus on extending the algorithm to multiple frames and to incorporate a generalized brightness model to deal with scenes with time-varying illumination.

Acknowledgment

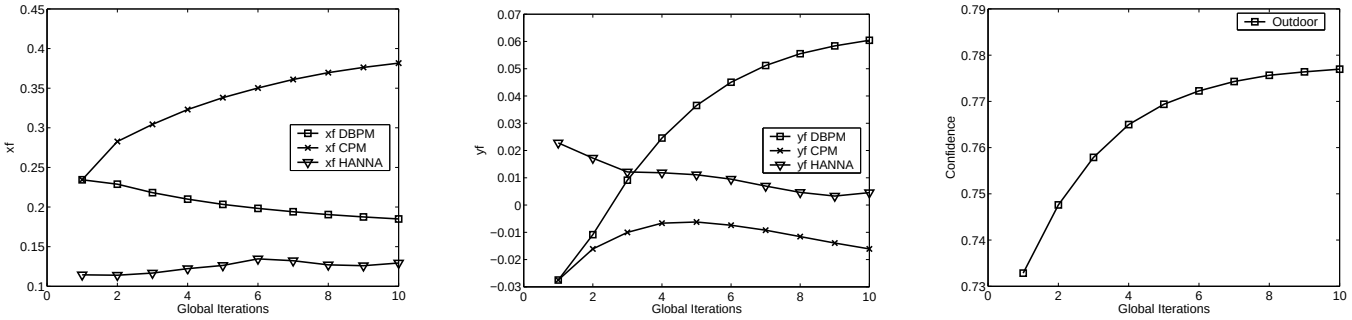
The authors would like to thank Phil David and Jeff DeHart of U.S. Army Research Laboratory for helpful discussions and



(a) Key image (b) Reference depth map (c) Texture mapped 3D model using DBPM

(d) Estimated depth map using DBPM (e) Estimated depth map using CPM (f) Estimated depth map using HANNA

Fig. 5. Outdoor



(a) (b) (c)

Fig. 6. Outdoor: (a and b) Convergence of x_f and y_f estimates. (c) Mean confidence over the entire image using DBPM.

assistance in data collection.

REFERENCES

[1] H. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133–135, 1981.

[2] T. Huang and A. Netravali, "Motion and structure from feature correspondences: A review," *Proc. IEEE*, vol. 82, pp. 252–268, Feb. 1994.

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

[4] J. Weng, T. Huang, and N. Ahuja, "3D motion estimation, understanding, and prediction from noisy image sequences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, pp. 370–389, 1987.

[5] J. Weng, N. Ahuja, and T. S. Huang, "Optimal motion and structure estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 864–884, Sept. 1993.

[6] A. Azarbayejani and A. Pentland, "Recursive estimation of motion, structure, and focal length," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 562–575, 1995.

[7] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int'l J. Computer Vision*, vol. 9, pp. 137–154, 1992.

[8] J. Oliensis, "A multi-frame structure-from-motion algorithm under perspective projection," *Int'l J. Computer Vision*, vol. 34, pp. 1–30, 1999.

[9] P. Beardsley, P. Torr, and A. Zisserman, "3D model acquisition from extended image sequences," in *Proc. European Conf. Computer Vision*, 1996, pp. 683–695.

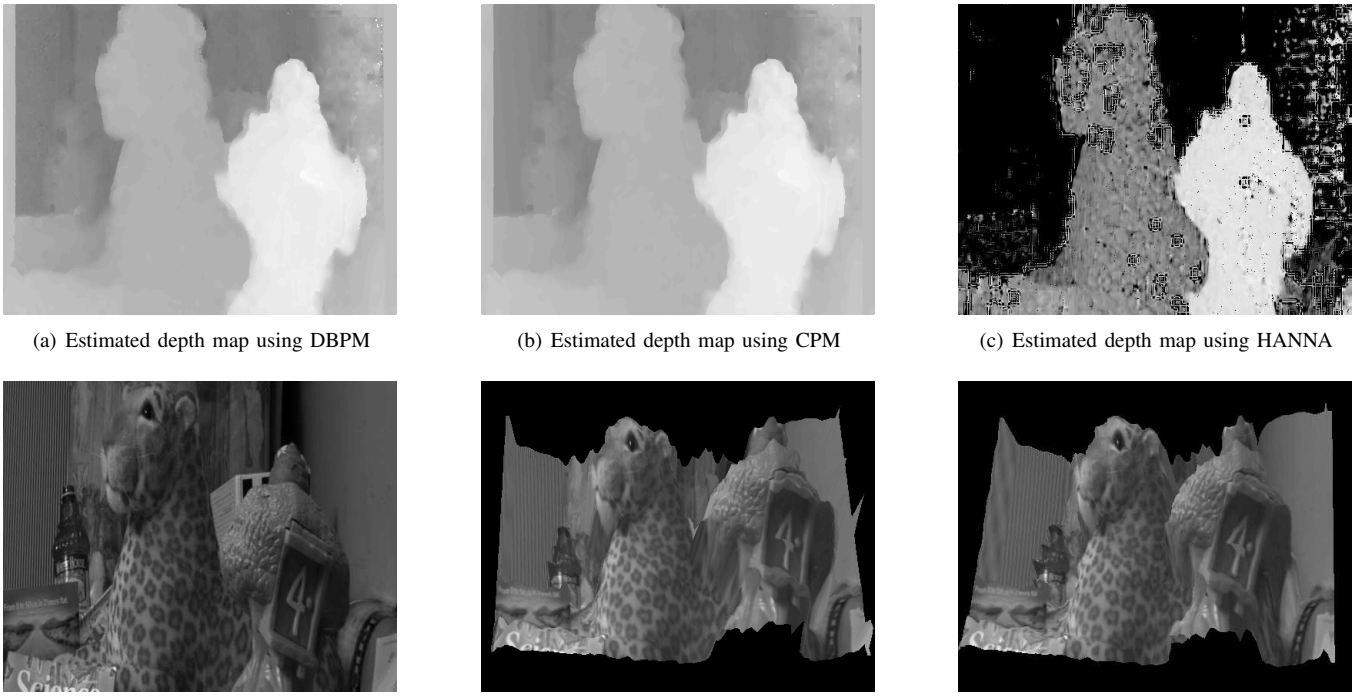
[10] H. Liu, R. Chellappa, and A. Rosenfeld, "A hierarchical approach for obtaining structure from two-frame optical flow," in *Proc. Workshop on Motion and Video Computing*, Orlando, FL, Dec. 2002, pp. 214–219.

[11] G. Adiv, "Determining 3D motion and structure from optical flow generated by several moving objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 7, no. 4, pp. 384–401, July 1985.

[12] A. Waxman and S. Ullman, "Surface structure and three-dimensional motion from image flow kinematics," *Int'l J. Robotics Research*, vol. 4, no. 3, pp. 72–94, 1985.

[13] R. Szeliski and S. Kang, "Recovering 3D shape and motion from image streams using non-linear least squares," *J. Visual Computation and Image Representation*, vol. 5, pp. 10–28, 1994.

[14] Y. Aloimonos and C. Brown, "Direct processing of curvilinear sensor



(a) Estimated depth map using DBPM (b) Estimated depth map using CPM (c) Estimated depth map using HANNA

(d) Key image (e) Novel View (f) Novel View

Fig. 8. LABS: (e and f) Rendered scene from novel viewpoints using depth map estimated using DBPM.

motion from a sequence of perspective images," in *Proc. IEEE Computer Society Workshop on Computer Vision: Representation and Control*, 1984, pp. 72–77.

[15] B. Horn and E. Weldon, "Direct methods for recovering motion," *Int'l J. Computer Vision*, pp. 51–76, 1988.

[16] K. Hanna, "Direct multi-resolution estimation of ego-motion and structure from motion," in *IEEE Workshop on Motion and Video Computing*, 1991, pp. 156–162.

[17] S. Negahdaripour, N. Kolagani, and B. Hayashi, "Direct motion stereo for passive navigation," in *Proc. Conf. Computer Vision and Pattern Recognition*, June 1992, pp. 425–431.

[18] R. Kumar, P. Anandan, and K. Hanna, "Direct recovery of shape from multiple views: a parallax based approach," in *Proc. 12th IAPR Int'l Conf. Pattern Recognition*, vol. 1, Jerusalem, Israel, 1994, pp. 685–688.

[19] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Proc. European Conf. Computer Vision*, Santa Margherita Ligure, Italy, May 1992, pp. 237–252.

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

[21] G. Stein and A. Shashua, "Model based brightness constraints: On direct estimation of structure and motion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 992–1015, Sept. 2000.

[22] S. Negahdaripour and J. Lanjing, "Direct recovery of motion and range from images of scenes with time-varying illumination," in *Proc. Int'l Symposium Computer Vision*, Nov. 1995, pp. 467–472.

[23] A. Shashua and N. Navab, "Relative affine structure: Theory and application to 3D reconstruction from perspective views," in *Proc. Conf. Computer Vision and Pattern Recognition*, June 1994, pp. 483–489.

[24] H. Sawhney, "3D geometry from planar parallax," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1994, pp. 929–934.

[25] M. Irani and P. Anandan, "Parallax geometry of pairs of points for 3D scene analysis," in *Proc. European Conf. Computer Vision*, Cambridge, UK, Apr. 1996, pp. 17–30.

[26] M. Irani, B. Rousso, and S. Peleg, "Recovery of ego-motion using region alignment," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 3, pp. 268–272, 1997.

[27] M. Irani, P. Anandan, and D. Weinshall, "From reference frames to reference planes: Multi-view parallax geometry and applications," in *Proc. European Conf. Computer Vision*, vol. II, June 1998, pp. 829–845.

[28] A. Dick and R. Cipolla, "Model refinement from planar parallax," in *Proc. 10th British Machine Vision Conf.*, Nottingham, UK, Sept. 1999, pp. 73–82.

[29] H. Jin, P. Favaro, and S. Soatto, "A semi-direct approach to structure from motion," in *The Visual Computer*, vol. 19, Oct. 2003, pp. 377–394.

[30] J. Kearney, W. Thompson, and D. Boley, "Optical flow estimation: An error analysis of gradient-based methods with local optimization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, pp. 229–244, Mar. 1987.

[31] T. Broida and R. Chellappa, "Performance bounds for estimating three-dimensional motion parameters from a sequence of noisy images," *J. Optical Society of America A*, vol. 6, pp. 879–889, 1989.

[32] G. S. Young and R. Chellappa, "Statistical analysis of inherent ambiguities in recovering 3D motion from a noisy flow field," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 995–1013, Oct. 1992.

[33] J. Oliensis, "The least-squares error for structure from infinitesimal motion," in *Proc. European Conf. Computer Vision*, 2004.

[34] A. Chiuso, R. Brockett, and S. Soatto, "Optimal structure from motion: Local ambiguities and global estimates," *Int'l J. Computer Vision*, vol. 39, no. 3, pp. 195–228, Sept. 2000.

[35] S. Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, ser. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, June 1991, vol. 9.

[36] H. Haussecker and D. Fleet, "Computing optical flow with physical models of brightness variation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 6, pp. 661–673, June 2001.

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

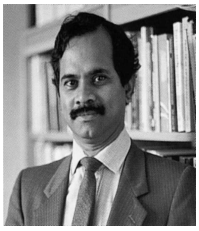
[38] J. Brandt, "Analysis of bias in gradient-based optical-flow estimation," in *IEEE Asilomar Conf. Signals, Systems and Computers*, Oct. 1995, pp. 721–725.

[39] H. Liu, T. Hong, M. Herman, and R. Chellappa, "A general motion model and spatio-temporal filters for computing optical flow," *Int'l J. Computer Vision*, vol. 22, pp. 141–172, 1997.

[40] B. Jahne, *Spatio-Temporal Image Processing, Theory and Scientific Applications*, ser. Lecture Notes in Computer Vision. Berlin, Germany: Springer-Verlag, 1993, vol. 751.

[41] ——, "Analytical studies of low-level motion estimators in space-time images using a unified filter concept," in *Proc. Conf. Computer Vision and Pattern Recognition*, June 1994, pp. 229–236.

Amit Agrawal received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 2000 and the M.S. degree in electrical and computer engineering from the University of Maryland, College Park, in 2003. From June 2000 to August 2001, he worked as a software engineer at the digital signal processing group at Hughes Software Systems, India, and Hughes Network Systems, MD, USA. He is currently a doctoral student in the Department of Electrical and Computer Engineering at the University of Maryland. His research interests are in computer vision, signal/image processing, biometrics and computer graphics.



Rama Chellappa received the B.E. (Hons.) degree from the University of Madras, India, in 1975 and the M.E. (Distinction) degree from the Indian Institute of Science, Bangalore, in 1977. He received the M.S.E.E. and Ph.D. Degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1978 and 1981 respectively. Since 1991, he has been a Professor of electrical engineering and an affiliate Professor of computer science at the University of Maryland, College Park. He is also affiliated with the Center for Automation Research (Director) and the Institute for Advanced Computer Studies (Permanent member). Prior to joining the University of Maryland, he was an Assistant (1981-1986) and Associate Professor (1986-1991) and Director of the Signal and Image Processing Institute (1988-1990) with the University of Southern California, Los Angeles. Over the last 23 years, he has published numerous book chapters, peer-reviewed journal and conference papers. He has edited a collection of Papers on Digital Image Processing (published by IEEE Computer Society Press), co-authored a research monograph on Artificial Neural Networks for Computer Vision (With Y.T. Zhou) published by Springer-Verlag, and co-edited a book on Markov Random fields (with A.K. Jain) published by Academic Press. His current research interests are face and gait analysis, 3D modeling from video, automatic target recognition from stationary and moving platforms, surveillance and monitoring, hyper spectral processing, image understanding, and commercial applications of image processing and understanding. Dr. Chellappa has served as an associate editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IMAGE PROCESSING, and NEURAL NETWORKS. He was co-Editor-in- Chief of Graphical models and Image Processing and served as the Editor-in-Chief of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE during 2001-2004. He also served as a member of the IEEE Signal Processing Society Board of Governors during 1996-1999 and was the Vice President of Awards and Membership during 2002-2004. He has received several awards, including NSF Presidential Young Investigator Award, an IBM Faculty Development Award, the 1990 Excellence in Teaching Award from School of Engineering at USC, the 1992 Best Industry Related Paper Award from the International Association of Pattern Recognition (with Q. Zheng) and the 2000 Technical Achievement Award from the IEEE Signal Processing Society. He was elected as a Distinguished Faculty Research Fellow (1996-1998) and as Distinguished Scholar-Teacher in 2003 at the University of Maryland. He is a Fellow of the International Association for Pattern Recognition. He has served as a General and Technical Program Chair for several IEEE international and national conferences and workshops.