3D MODEL REFINEMENT USING SURFACE-PARALLAX

Amit K Agrawal and Rama Chellappa

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

ABSTRACT

We present an approach to update and refine coarse 3D models of urban environments from a sequence of intensity images using surface parallax. This generalizes the plane-parallax recovery methods to surface-parallax using arbitrary surfaces. A coarse and potentially incomplete depth map of the scene obtained from a Digital Elevation Map (DEM) is used as a reference surface which is refined and updated using this approach. The reference depth map is used to estimate the camera motion and the motion of the 3D points on the reference surface is compensated. The resulting parallax, which is an epipolar field, is estimated using an adaptive windowing technique and used to obtain the refined depth map.

1. INTRODUCTION

3D reconstruction from images or multiple views has been an active research area over the past few decades. Traditional methods have focused on expressing the image motion of rigid objects as a sum of translational and rotational fields. Alternative approaches based on decomposing the image motion due to a parametric surface and a residual parallax field have also been proposed. Most of the previous methods [1][2][3] either focus on using a planar surface as the surface for alignment and assume the presence of a dominant planar surface in the scene or work well on piecewise planar models. These methods first identify a dominant planar surface in the scene and then use it to warp the images to a reference frame. The resulting parallax is due to static scene points not on the planar surface or due to independently moving objects. Assuming no independent movements, parallax can be used to estimate the depth of points not on the planar surface. In [4], the authors present the *Facade* system which can produce highly realistic models of architectural scenes from photographs. However, it requires users to specify a polyhedral model and is particularly suited for architectural scenes.

In this work, we extend the ”plane+parallax” approach to a ”surface+parallax” approach in which we use an arbitrary surface for alignment and hence do not require assumptions such as dominant planar surfaces be present in the scene. This approach is partially facilitated by the availability of Digital Elevation Maps (DEM) of urban environments. In general, these DEM's are coarse (low resolution) and may contain partial information about the area due to topographical and structural changes (e.g. construction, demolition of buildings). They can, however be used to obtain a coarse reference surface for the scene which can be updated and refined using information from a sequence of images of the scene. The only assumption we make is that the scene contains a small planar surface which is used to estimate the camera motion. This planar surface can be as small as occupying only $5-10$ percent of image area and is used only for camera motion estimation. Apart from that, the reference surface can be as arbitrary as possible and the parallax is estimated by aligning the entire reference surface.

The advantages of using this approach over plane-parallax approach are as follows.

- Using this approach, more complex scenes can be handled since the entire reference surface is used for alignment. The assumption of a small planar surface for camera motion estimation is not restrictive and can be easily met in outdoor urban environments.
- In this approach, we first estimate the camera motion and hence the Focus of Expansion (FOE). Since the parallax field is an epipolar field, for each pixel we know the parallax direction from the FOE. The estimation of parallax magnitude (as shown in later sections) is then a simple linear problem for each pixel which can be solved using least squares or total least squares. However, for the other methods, one has to estimate the FOE (or equivalently parallax direction) along with parallax magnitude which is more difficult.

These advantages are due to the fact that we are utilizing prior information in the form of a coarse depth map (reference surface).

The rest of the paper is organized as follows. In the next section, we explain the theory behind surface parallax. Section 3 explains our algorithm followed by experiments and conclusions at the end.

2. SURFACE PARALLAX

The principle behind surface parallax is that for any two views of a scene under perspective projection, if the motion of the 3D points on a surface is compensated, the resulting parallax field is an epipolar field. Referring to Figure 1, let C_1 and C_2 represent the camera center for two views and S be the reference surface which is aligned. Let Q be the 3D point on the reference surface, P be the true location of the 3D point and the projection of these points in reference image C_1 be q and p respectively. The residual

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.

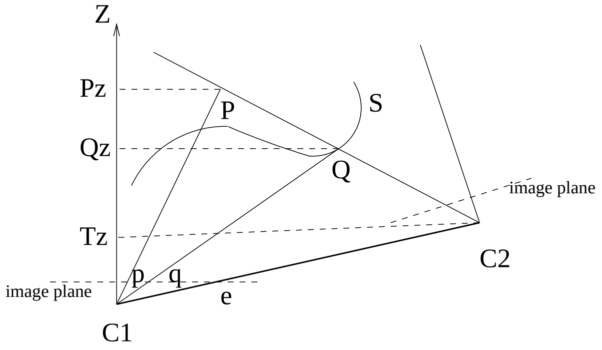

Fig. 1. Parallax due to surface S

parallax can be shown to be equal to [1]

$$\delta u = q - p = \frac{T_z(Q_z - P_z)}{Q_z(P_z - T_z)}(p - e) \tag{1}$$

where e denotes the epipole and T_z denotes the translation in Z direction. If $T_z = 0$,

$$\delta u = q - p = \frac{-f(Q_z - P_z)}{Q_z P_z}(t) \tag{2}$$

where f is the focal length and $t = [T_x, Ty]^T$ denotes the 2×1 translation vector in x, y space.

Since (1) has the unknown correspondence p on right hand side, it is solved for parallax in terms of q as

$$\delta u = q - p = \frac{T_z(Q_z - P_z)}{P_z(Q_z - T_z)}(q - e) \tag{3}$$

Then $\beta = \frac{T_z(Q_z - P_z)}{P_z(Q_z - T_z)}\|(q - e)\|$ denotes the parallax magnitude and **v** $= \frac{(q-e)}{\|(q-e)\|}$ denotes the parallax direction. For the case when $T_z = 0$, we have $\beta = \frac{-f(Q_z - P_z)}{Q_z P_z}\sqrt[2]{T_x^2 + T_y^2}$ and parallax direction **v** $= \frac{(Tx, Ty)^T}{\sqrt[2]{T_x^2 + T_y^2}}$ is constant for all pixels.

3. ALGORITHM

Our approach aligns an arbitrary non-planar surface (hereby referred to as the *reference surface*) in images and estimate the deviations from the reference surface by calculating the residual parallax field obtained after the alignment. The algorithm uses two frames from the image sequence, one of them being the reference frame for which the depth map is refined. The reference surface is rendered to give the depth map for the reference image. So as an input to the algorithm, we have a reference image, a second frame from the sequence and the depth map for the reference image (coarse and potentially incorrect) which we wish to refine and update.

3.1. Coordinate System

For the rest of the paper, we assume a camera centered coordinate system with origin at the center of the reference frame camera. The Z axis points along the principal axis of the camera with negative Z in front of the camera (in accordance with OpenGL conventions).

3.2. Estimating Camera Motion

We begin by first estimating the camera motion using an optical flow based structure from motion algorithm. Consider the equations relating the image motion of a rigid body with depth and camera motion [5]

$$\begin{aligned} u(x, y) = {} & \frac{-fT_x + xT_z}{Z} \\ & + \frac{1}{f}xy\Omega_x - (f + \frac{1}{f}x^2)\Omega_y + y\Omega_z \end{aligned} \tag{4}$$

$$\begin{aligned} v(x, y) = {} & \frac{-fT_y + yT_z}{Z} \\ & + (f + \frac{1}{f}y^2)\Omega_x - \frac{1}{f}xy\Omega_y - x\Omega_z \end{aligned} \tag{5}$$

where $(u, v)^T$ denotes the 2-D velocities, Z is the depth and $(T_x, T_y, T_z)^T$ and $(\Omega_x, \Omega_y, \Omega_z)^T$ denote the camera translation and rotation velocities respectively. It is well known that the above system is a bilinear system. If 2-D velocities and depth are known, the system is linear in camera motion and if 2-D velocities and motion is known, the system is linear in depth.

We first identify a small planar region in the 3D scene and its corresponding region in the reference image. Since the optical flow of a planar surface is parametric, we fit a parametric optical flow [6] to the region and obtain (u, v) for that region. Since we know the coarse depth, we can estimate the motion using the above equations and then use the estimated motion to refine the depth map [7]. This can be iterated until the motion estimate is stable or a specified number of iterations are reached.

3.3. Aligning the Reference Surface

Let $(u(x, y), v(x, y))$ denotes the true optical flow of pixel (x, y) in the reference frame. It can be decomposed as

$$u(x, y) = u_{Z_{ref}}(x, y) + u_p(x, y) \tag{6}$$

$$v(x, y) = v_{Z_{ref}}(x, y) + v_p(x, y) \tag{7}$$

where $(u_{Z_{ref}}, v_{Z_{ref}})$ denotes the flow due to reference surface Z_{ref} and (u_p, v_p) denotes the parallax due to Z_{ref}. From brightness constancy, we have

$$I(x, y, t) = I(x - u(x, y), y - v(x, y), t - 1) \tag{8}$$

Let $I_1 = I(x, y, t)$ denotes the reference frame and $I_2 = I(x, y, t-1)$ denotes the previous frame in the sequence. Assuming a small parallax field, we make the approximation

$$\begin{aligned} I_1(x + u_p(x, y), y + v_p(x, y)) = I_2(x - u_{Z_{ref}}(x, y), \\ y - v_{Z_{ref}}(x, y)) \end{aligned} \tag{9}$$

Expanding the left hand side of the above equation in Taylor series around (x, y) and neglecting higher order terms, we have,

$$I_x u_p + I_y v_p + \Delta I = 0 \tag{10}$$

where I_x and I_y denotes the spatial image gradients and $\Delta I = I_1(x, y) - I_2(x - u_{Z_{ref}}(x, y), y - v_{Z_{ref}}(x, y))$ represents the difference between the key image and the *warped* offset image according to Z_{ref} and motion estimates. $(u_{Z_{ref}}, v_{Z_{ref}})$ is calculated from Z_{ref} and motion estimates using (4) and (5) and the previous image is warped towards reference image using bilinear interpolation.

3.4. Estimation of Parallax Field

Equation (10) shows the relationship between the image derivatives, parallax and the difference between the reference image and the warped previous image according to reference Z. Since we know the camera motion and hence the FOE (defined as (x_f, y_f)), we can write the parallax field as

$$u_p(x,y) = \beta(x,y)du(x,y) \tag{11}$$

$$v_p(x,y) = \beta(x,y)dv(x,y) \tag{12}$$

where $du(x,y) = \frac{(x-x_f)}{\sqrt[2]{(x-x_f)^2+(y-y_f)^2}}$ and $dv(x,y) = \frac{(y-y_f)}{\sqrt[2]{(x-x_f)^2+(y-y_f)^2}}$. In the above equation, $\beta(x,y)$ denotes the parallax magnitude and $(du(x,y), dv(x,y))$ denotes the parallax direction for pixel (x,y). Equation (10) then becomes

$$\beta(x,y)(I_x du(x,y) + I_y dv(x,y)) + \Delta I(x,y) = 0 \tag{13}$$

which is a linear system for each pixel (x,y).

We cannot estimate the parallax magnitude for each pixel using (13) because it is very sensitive to noise. We need to regularize the solution. We assume that the parallax magnitude is constant over a neighborhood $N \times N$. Thus for pixel $(\overline{x},\overline{y})$ we minimize the following error function in least square sense

$$\begin{aligned} J(\overline{x},\overline{y}) = \min_{\alpha} & \sum_{(x,y)\in N\times N} w(x-\overline{x}, y-\overline{y}) \\ & \times \{\alpha(I_x du(x,y) + I_y dv(x,y)) \\ & + \Delta I(x,y)\}^2 \end{aligned} \tag{14}$$

where $w(x,y)$ is a window function centered around $(\overline{x},\overline{y})$. This leads to a least square solution for the estimated parallax magnitude $\beta(\overline{x},\overline{y}) = \alpha$.

We can also formulate the problem as a total least square solution. Consider minimizing the following error function

$$\begin{aligned} J(\overline{x},\overline{y}) = \min_{\mu,\nu} & \sum_{(x,y)\in N\times N} w(x-\overline{x}, y-\overline{y}) \\ & \times \{\mu(I_x du(x,y) + I_y dv(x,y)) \\ & + \nu\Delta I(x,y)\}^2 \end{aligned}$$

subject to

$$\mu^2 + \nu^2 = 1 \tag{15}$$

Then the parallax magnitude will be given by

$$\beta(\overline{x},\overline{y}) = \frac{\mu}{\nu} \tag{16}$$

We use an adaptive windowing based method [8] for estimating the parallax using the total least squares method.

3.5. Depth Refinement and Update using Parallax

From (3), the true depth P_z can be solved in terms of reference depth Q_z and parallax as (for $T_z \neq 0$)

$$\begin{aligned} \gamma(x,y) &= \frac{\beta(x,y)}{\|(q-e)\|} \\ &= \frac{\beta(x,y)}{\sqrt[2]{(x-x_f)^2+(y-y_f)^2}} \end{aligned} \tag{17}$$

$$P_z(x,y) = \frac{T_z Q_z(x,y)}{\gamma(x,y)(Q_z(x,y)-T_z)+T_z} \tag{18}$$

For $T_z = 0$,

$$\gamma(x,y) = \frac{\beta(x,y)}{-f\sqrt[2]{Tx^2+Ty^2}} \tag{19}$$

$$P_z(x,y) = \frac{Q_z(x,y)}{\gamma(x,y)Q_z(x,y)+1} \tag{20}$$

4. EXPERIMENTS

We present results using a semi-synthetic 3D model with real textures. The 3D model is rendered in OpenGL. We simulate a sequence of images by moving a virtual camera in the scene. The reference coarse depth map is obtained using OpenGL Z buffer. In real scenarios, the reference depth map can be obtained by rendering the available DEM in OpenGL. Since the textures are real, we face the problems related to optical flow estimation normally encountered in real world images.

The image sequence was filtered using a gaussian filter with $\sigma = 2.5$ for temporal axis and $\sigma = 1$ for spatial axis. Figure 2 shows two images from the sequence used for parallax computation. The camera is translating and rotating in this sequence. A portion of the ground plane was used for camera motion estimation. The true and estimated camera motions are given in Table 1.

	Tx	Ty	Tz	Wx	Wy	Wz
I	-2.505	0.239	2.484	0.50	0.01	0.50
II	-2.481	0.150	2.330	0.49	-0.02	0.52

Table 1. True (I) and estimated (II) motion parameters

The reference depth map which is used as the surface for alignment is shown in Figure 3. The surface (other than the ground plane) in Figure 3 is rendered in OpenGL using the equation $Z = -300 - 100 \times sin(\pi * (X+300)/600)$ where (X,Y,Z) denotes a 3D point on the surface. Thus, the reference surface is highly non-planar. The depth map was refined using the total least square solution. Homogeneous regions were identified by thresholding the magnitude of intensity gradient and regions where local edge structure is aligned in the direction of FOE (i.e. $I_x du + I_y dv \approx 0$) were identified by thresholding the magnitude of $(I_x du + I_y dv)$. For these regions, parallax can not be computed reliably. The thresholds for above two cases was set to 0.1.

Figure 4 shows the true and estimated depth maps. The true depth map has several objects in front whose depths are estimated using our approach. We define the relative mean square error (RMSE) between true depth map Z_{true} and some other depth map Z as

$$RMSE = 100 \times \frac{1}{N}\sum_{1}^{N}(\frac{Z_{true}-Z}{Z_{true}})^2 \tag{21}$$

where N denotes the total number of pixels in the image. For this example, $RMSE = 36.73\%$ between true and reference depth map. Final relative mean square error between the true and estimated depth maps is $RMSE = 8.50\%$ which shows the improvement in the depth map using our approach.

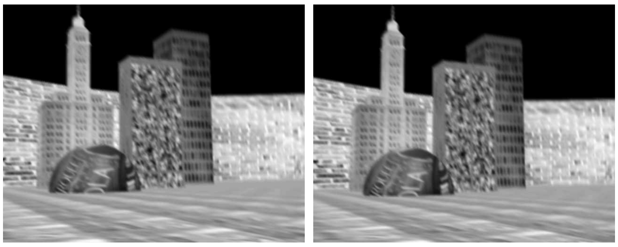

Fig. 2. Reference frame (a) and previous frame (b) for translating and rotating camera motion

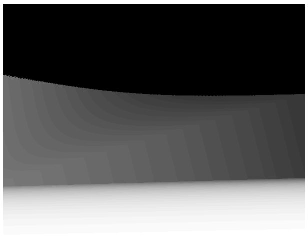

Fig. 3. Reference depth map

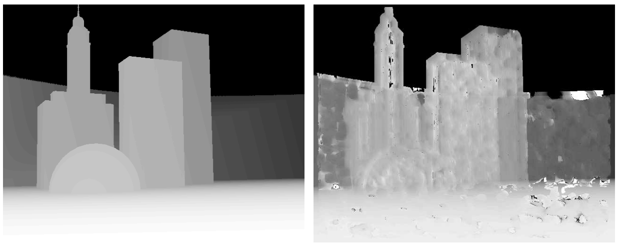

Fig. 4. (a) True depth map (b) Estimated depth map

5. CONCLUSIONS

In this paper, we have presented an approach for 3D model refinement given a coarse/incomplete depth map and intensity images using surface parallax. This can be viewed as an extension to plane-parallax approaches. Current results on semi-synthetic sequences shows the validity of our approach. Future research will focus on extending the algorithm to more than two frames.

6. REFERENCES

[1] R. Kumar, P. Anandan, and K. 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.

[2] A.R. Dick and R. Cipolla, “Model refinement from planar parallax,” *Proc. 10th British Machine Vision Conference*, pp. 73–82, 1999.

[3] 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.

[4] C.J. Taylor, P.E. Debevec, and J. Malik, “Modeling and rendering architecture from photographs: A hybrid geometry and image-based approach,” *Technical report, University of California at Berkeley*, 1996.

[5] B.K.P. Horn, “Robot vision,” *McGraw-Hill*, 1986.

[6] 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.

[7] 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.

[8] 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.