Dense Depth Maps from Epipolar Images

J.P. Mellor * **Seth Teller** **Tomás Lozano-Pérez**
MIT Artificial Intelligence Laboratory
MIT Computer Graphics Laboratory
545 Technology Square NE43-753
Cambridge, MA 02139
jpmellor@ai.mit.edu, http://www.graphics.lcs.mit.edu

Abstract

This paper describes a method for generating dense depth maps given large numbers of images taken from arbitrary positions. The algorithm presented is completely local and uses an epipolar image to generate for each pixel an evidence versus depth and surface normal distribution. In many cases, the distribution contains a clear and distinct global maximum. The location of this peak determines the depth and its shape can be used to estimate the error. The distribution can also be used to perform a maximum likelihood fit of models directly to the images. We anticipate that the ability to perform maximum likelihood estimation from purely local calculations will prove useful in constructing three dimensional models from large sets of images.

1 Introduction

One approach to improving the results obtained by stereo techniques is to use multiple images. Several researchers, such as Yachida [1986], have proposed trinocular stereo algorithms. Others have also used special camera configurations to aid in the correspondence problem [Tsai, 1983, Bolles *et al.*, 1987, Okutomi and Kanade, 1993]. The work presented here also uses multiple images and draws its major inspiration from Bolles, Baker and Marimont [1987]. We define a construct called an *epipolar image* and use it to analyze evidence about depth. Like Tsai [1983] and Okutomi and Kanade [1993] we define a cost function that is applied across multiple images, and like Cox [1996] we model the occlusion process. There are several important differences, however. The epipolar image we define is valid for arbitrary camera positions and models some forms of occlusion. Our method is intended to recover dense depth maps of built geometry (architectural facades) using thousands of images acquired from within the scene. In most cases, depth can be recovered using purely local information, avoiding the computational costs of global constraints. Where depth cannot be recovered using purely local information, the depth evidence from the epipolar image provides a principled distribution for use in a maximum-likelihood approach [Duda and Hart, 1973].

2 Our Approach

We assume that camera pose is known in an absolute coordinate system. Although relative positions are sufficient for the discussion in this section, global positions allow us to

*This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of the Defense under Office of Naval Research contract N00014-91-J-4038. The author was also supported by the Advanced Research Projects Agency of the Department of Defense under Rome Laboratory contract F3060-94-C-0204.

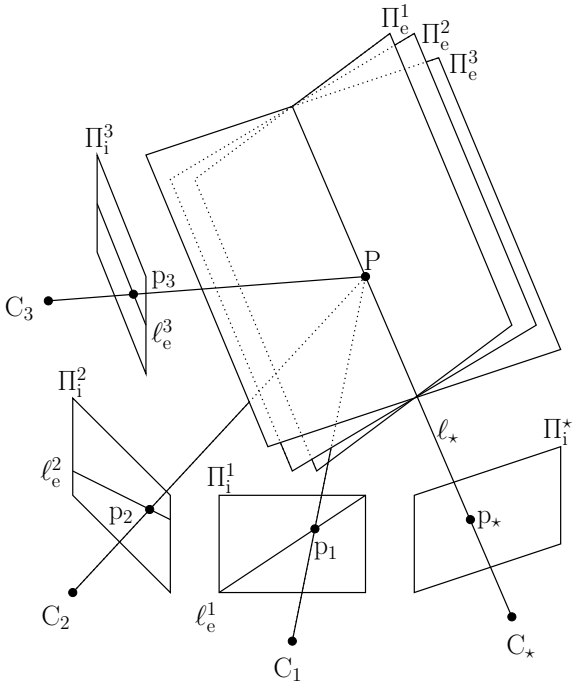

Figure 1: Epipolar image geometry.

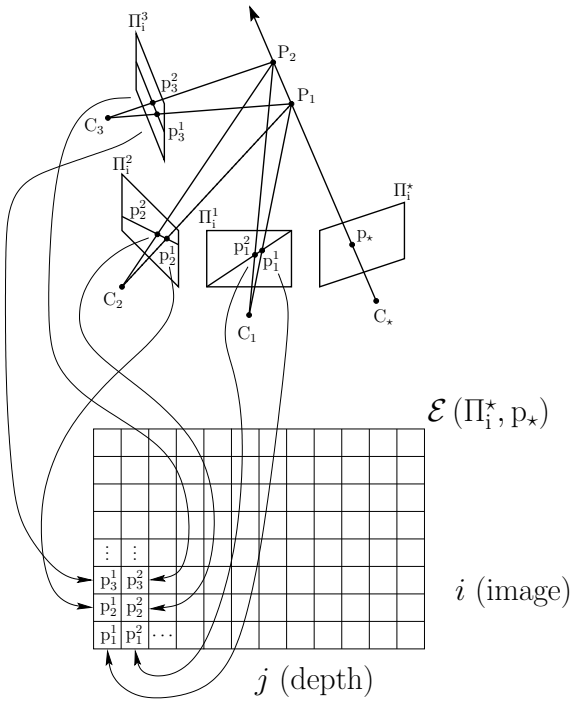

Figure 2: Constructing an epipolar image.

perform reconstruction incrementally using disjoint scenes. We also assume known internal camera parameters. For a more complete description of our method see [Mellor *et al.*, 1996].

2.1 Epipolar Images

For our analysis we will define an epipolar image $\mathcal{E}$ which is a function of one image and a point in that image. An epipolar image is similar to an epipolar-plane image [Bolles *et al.*, 1987], but has one critical difference that ensures it can be constructed for *every* pixel in an arbitrary set of images. Rather than use projections of a single epipolar plane, we construct the epipolar image from the *pencil* of epipolar planes defined by the line $\ell_\star$ through one of the camera centers $C_\star$ and one of the pixels $p_\star$ in that image[1] $\Pi_i^\star$ (Figure 1). Π_e^i is the epipolar plane formed by $\ell_\star$ and the i^{th} camera center C_i. Epipolar line ℓ_e^i contains all of the information about $\ell_\star$ present in Π_i^i.

To simplify the analysis of an epipolar image we can group points from the epipolar lines according to possible correspondences (Figure 2). P_1 projects to p_i^1 in Π_i^i; therefore $\{p_i^1\}$ has all of the information contained in $\{\Pi_i^i\}$ about P_1. Similarly, there is a distinct set for P_2; thus $\left\{p_i^j \,|\, \text{for a given } j\right\}$ contains all of the possible correspondences for P_j. If P_j is a point on the surface of a physical object and it is visible in $\{\Pi_i^i\}$ and $\Pi_i^\star$, then measurements taken at p_i^j should match[2] those taken at $p_\star$ (Figure 3a). Conversely, if P_j is not a point on the surface of a physical object then the measurements taken at p_i^j are unlikely to match those taken at $p_\star$ (Figures 3b and 3c). Epipolar images can be viewed as tables which accumulate evidence about possible correspondences of $p_\star$. A simple function of j is used to build $\left\{P_j \,|\, \forall i < j : \|P_i - C_\star\|^2 < \|P_j - C_\star\|^2\right\}$. In essence, $\{P_j\}$ is a set of samples along $\ell_\star$ at increasing depths from the image plane of $p_\star$.

[1] Π is used to denote a plane; the subscript identifies the type (**e**pipolar or **i**mage); and the superscript identifies the instance.

[2] So far we have considered only diffuse surfaces. The matching function can be extended to account for specularity and we intend to do so in the future.

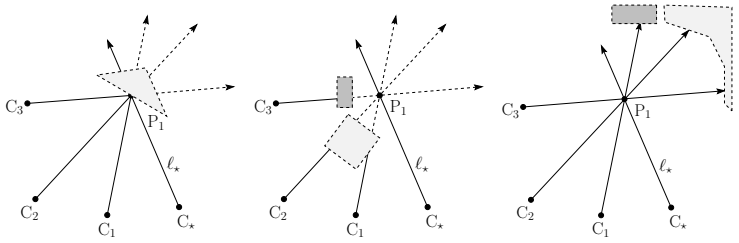

Figure 3: Typical cases.

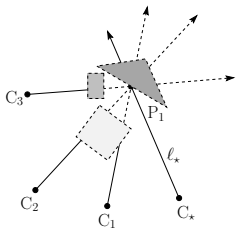

Figure 4: False negative.

2.2 Analyzing Epipolar Images

An epipolar image $\mathcal{E}$ is constructed by organizing image measurements into a two-dimensional array with i and j as the vertical and horizontal axes respectively. Rows in $\mathcal{E}$ are epipolar lines from different images; columns form sets of possible correspondences ordered by depth[3] (Figure 2). The quality $\nu(j)$ of the match between column j and $\mathrm{p}_\star$ can be thought of as evidence that $\mathrm{p}_\star$ is the projection of P_j with depth j. Real cameras have a finite field of view, and p_i^j may not be contained in the image Π_{i}^i $\left(\mathrm{p}_i^j \notin \{\Pi_{\mathrm{i}}^i\}\right)$. Thus, only terms for which $\mathrm{p}_i^j \in \{\Pi_{\mathrm{i}}^i\}$ should be included, producing

$$\nu(j) = \frac{\sum_{i \,|\, \mathrm{p}_i^j \in \{\Pi_{\mathrm{i}}^i\}} \mathcal{X}(\mathcal{F}(\mathrm{p}_i^j), \mathcal{F}(\mathrm{p}_\star))}{\sum_{i \,|\, \mathrm{p}_i^j \in \{\Pi_{\mathrm{i}}^i\}} 1} \qquad (1)$$

where $\mathcal{F}()$ is an image measurement and $\mathcal{X}()$ is a cost function which measures the difference between $\mathcal{F}(\mathrm{p}_i^j)$ and $\mathcal{F}(\mathrm{p}_\star)$. Ideally, $\nu(j)$ will have a sharp, distinct peak at the correct depth, so that

$$\arg\max_j(\nu(j)) = \text{ the correct depth of } \mathrm{p}_\star.$$

As the number of elements in $\left\{\mathrm{p}_i^j \,|\, \text{for a given } j\right\}$ increases, the likelihood increases that $\nu(j)$ will be large when P_j lies on a physical surface and small when it does not. Occlusion does not produce a peak at an incorrect depth or a false positive[4]. It can however, cause false negatives or the absence of a peak at the correct depth (Figure 4). A false negative is essentially a lack of evidence about the correct depth and can be addressed in two ways: removing the contribution of occluded views, and adding unoccluded views by acquiring more images.

A large class of occluded views can be eliminated quite simply. Each point P_j has an associated normal n_j. Images with camera centers in the negative half space defined by n_j cannot possibly have imaged P_j. Of course, n_j is not known a priori, but the fact that P_j is visible in $\Pi_{\mathrm{i}}^\star$ limits its possible values. This range of values can then be sampled and used to eliminate the contribution of occluded views from $\nu(j)$. Let α be an estimate of n_j and $\widehat{\mathrm{C}_i\mathrm{P}_j}$ be the unit vector along the ray from C_i to P_j, then P_j can only be visible if $\widehat{\mathrm{C}_i\mathrm{P}_j} \cdot \alpha < 0$. The updated function becomes:

$$\nu(j,\alpha) = \frac{\sum_{i \in \mathcal{S}} \left(\widehat{\mathrm{C}_i\mathrm{P}_j} \cdot \alpha\right) \mathcal{X}(\mathcal{F}(\mathrm{p}_i^j), \mathcal{F}(\mathrm{p}_\star))}{\sum_{i \in \mathcal{S}} \widehat{\mathrm{C}_i\mathrm{P}_j} \cdot \alpha} \qquad (2)$$

where

$$\mathcal{S} = \left\{ i \;\middle|\; \begin{array}{l} \mathrm{p}_i^j \in \{\Pi_{\mathrm{i}}^i\} \\ \widehat{\mathrm{C}_i\mathrm{P}_j} \cdot \alpha < 0 \end{array} \right\}.$$

Then, if sufficient evidence exists,

$$\arg\max_{j,\alpha}(\nu(j,\alpha)) \;\Rightarrow\; \left\{ \begin{array}{l} j = \text{ depth of } \mathrm{p}_\star \\ \alpha \text{ an estimate of } \mathrm{n}_j \end{array} \right. .$$

3 Results

Synthetic imagery was used to explore the characteristics of $\nu(j)$ and $\nu(j,\alpha)$. A CAD model of Technology Square, the four-building complex housing our laboratory, was

[3] The depth of P_j can be trivially calculated from j, therefore we consider j and depth to be interchangeable.

[4] Except possibly in an adversarial setting.

built by hand. The locations and geometries of the buildings were determined using traditional survey techniques. Photographs of the buildings were used to extract texture maps which were matched with the survey data. This three-dimensional model was then rendered from 100 positions along a "walk around the block". From this set of images, a $\Pi_{\mathrm{i}}^{\star}$ and $\mathrm{p}_{\star}$ were chosen and an epipolar image $\mathcal{E}$ constructed. $\mathcal{E}$ was then analyzed using equations 1 and 2 where[5]

$$\mathcal{F}(x) = \mathrm{hsv}(x) = [\mathrm{h}(x), \mathrm{s}(x), \mathrm{v}(x)]^{\mathrm{T}} \quad (3)$$

and

$$\begin{aligned} \mathcal{X}([h_1, s_1, v_1]^{\mathrm{T}}, [h_2, s_2, v_2]^{\mathrm{T}}) = \quad (4) \\ -\left(\frac{s_1 + s_2}{2}\right)(1 - \cos(h_1 - h_2)) - \\ (2 - s_1 - s_2)\,|v_1 - v_2|\,. \end{aligned}$$

Figure 5 shows a base image $\Pi_{\mathrm{i}}^{\star}$ with $\mathrm{p}_{\star}$ marked by a cross. Under $\Pi_{\mathrm{i}}^{\star}$ is the epipolar image $\mathcal{E}$ generated using the remaining 99 images. Below $\mathcal{E}$ is the matching function $\nu(j)$ (1) and $\nu(j,\alpha)$ (2). The horizontal scale, j or depth, is the same for $\mathcal{E}$, $\nu(j)$ and $\nu(j,\alpha)$. The vertical axis of $\mathcal{E}$ is the image index, and of $\nu(j,\alpha)$ is a coarse estimate of the orientation α at P_j. The vertical axis of $\nu(j)$ has no significance; it is a single row that has been replicated for clarity. To the right, $\nu(j)$ and $\nu(j,\alpha)$ are also shown as two-dimensional plots[6].

Figure 5a shows the epipolar image that results when the upper left-hand corner of the foreground building is chosen as $\mathrm{p}_{\star}$. Near the bottom of $\mathcal{E}$, ℓ_{e}^{i} is close to horizontal, and p_i^j is the projection of blue sky everywhere except at the building corner. The corner points show up in $\mathcal{E}$ near the right side as a vertical streak. This is as expected since the construction of $\mathcal{E}$ places the projections of P_j in the same column. Near the middle of $\mathcal{E}$, the long horizontal streaks result because P_j is occluded, and near the top the large black region is produced because $\mathrm{p}_i^j \notin \Pi_{\mathrm{i}}^{i}$. Both $\nu(j)$ and $\nu(j,\alpha)$ have a sharp peak[7] that corresponds to the vertical stack of corner points. This peak occurs at a depth of 2375 units ($j = 321$) for $\nu(j)$ and a depth of 2385 ($j = 322$) for $\nu(j,\alpha)$. The actual distance to the corner is 2387.4 units. The reconstructed world coordinates of $\mathrm{p}_{\star}$ are $[-1441, -3084, 1830]^{\mathrm{T}}$ and $[-1438, -3077, 1837]^{\mathrm{T}}$ respectively. The actual coordinates[8] are $[-1446, -3078, 1846]^{\mathrm{T}}$.

In Figure 5b, $\mathrm{p}_{\star}$ is a point from the interior of a building face with highly periodic texture. There is a clear peak in $\nu(j,\alpha)$ that agrees well with manual measurements and is better than that in $\nu(j)$. In Figure 5c, $\mathrm{p}_{\star}$ is a point on a building face that is occluded (Figure 4) in a number of views. Both $\nu(j)$ and $\nu(j,\alpha)$ produce fairly good peaks that agree with manual measurements.

To further test our method, we reconstructed the depth of a region in one of the images (Figure 6). For each pixel inside the black rectangle the global maximum of $\nu(j,\alpha)$ was taken as the depth of that pixel. Figure 7 shows the reconstructed world coordinates[9] for each of the 3000 pixels in the region. The cluster of points beyond the left end (near [0,2]) and at the right end of the building correspond to sky points. The actual world coordinates were calculated from the CAD model and are shown in grey. The camera position is marked by a grey line extending from the center of projection in the direction of the optical axis. The reconstruction was performed purely locally at each pixel. Global constraints such as ordering or smoothness were not imposed, and no attempt was made to remove depths with low confidence or otherwise post-process the global maximum of $\nu(j,\alpha)$.

[5] The well known hue, saturation, value color model is denoted by hsv.

[6] Actually, $\sum_{\alpha} \nu(j,\alpha) / \sum_{\alpha} 1$ is plotted for $\nu(j,\alpha)$.

[7] White indicates minimum error, black maximum.

[8] Some of the difference may be due to the fact that $\mathrm{p}_{\star}$ was chosen by hand and might not be the exact projection of the corner.

[9] All coordinates have been divided by 1000 to simplify the plots.

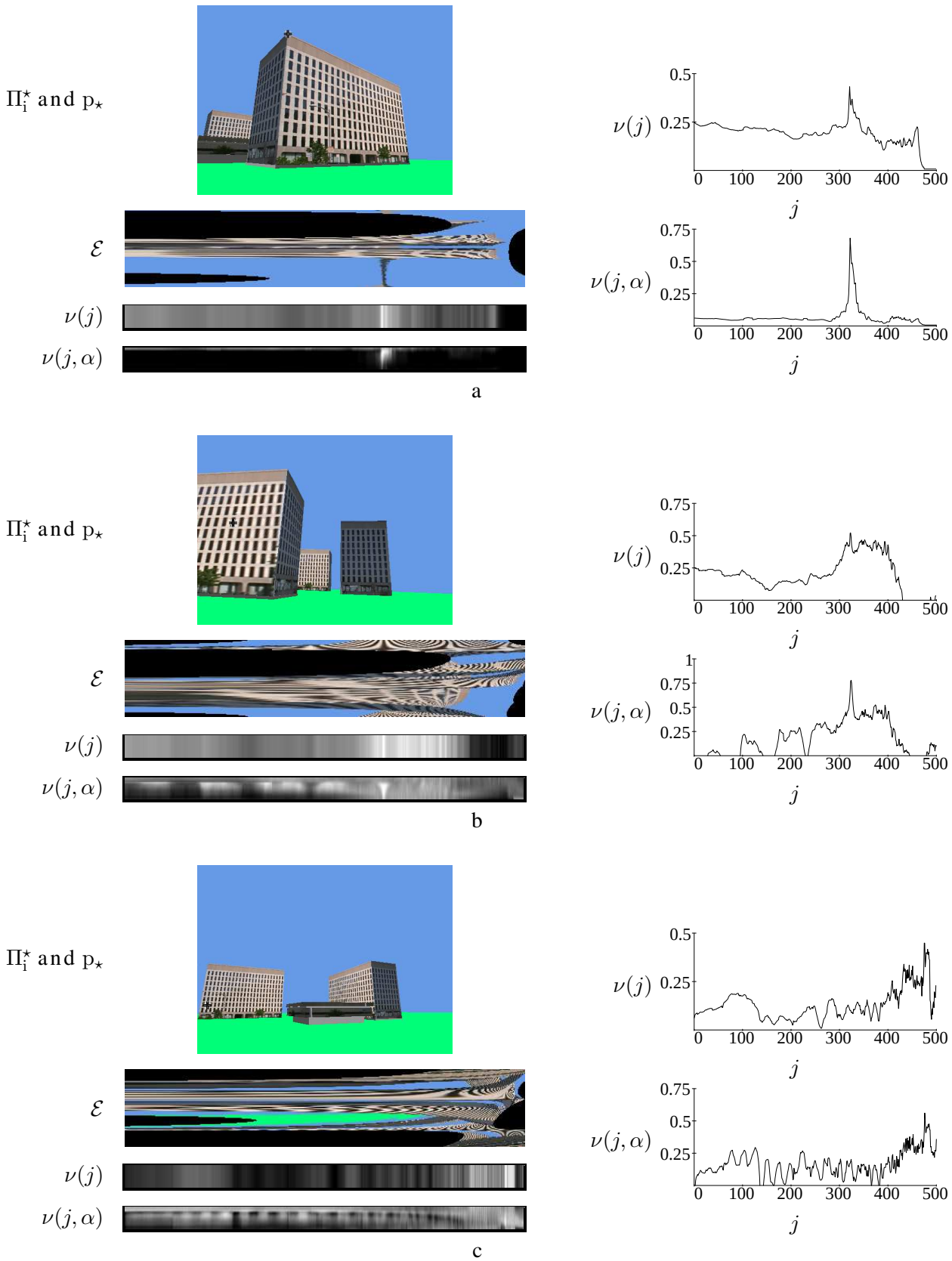

Figure 5: $\Pi_i^\star$, $p_\star$, $\mathcal{E}$, $\nu(j)$ and $\nu(j,\alpha)$.

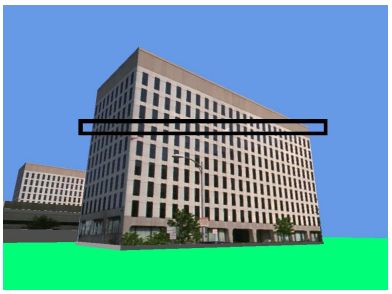

Figure 6: Reconstructed region.

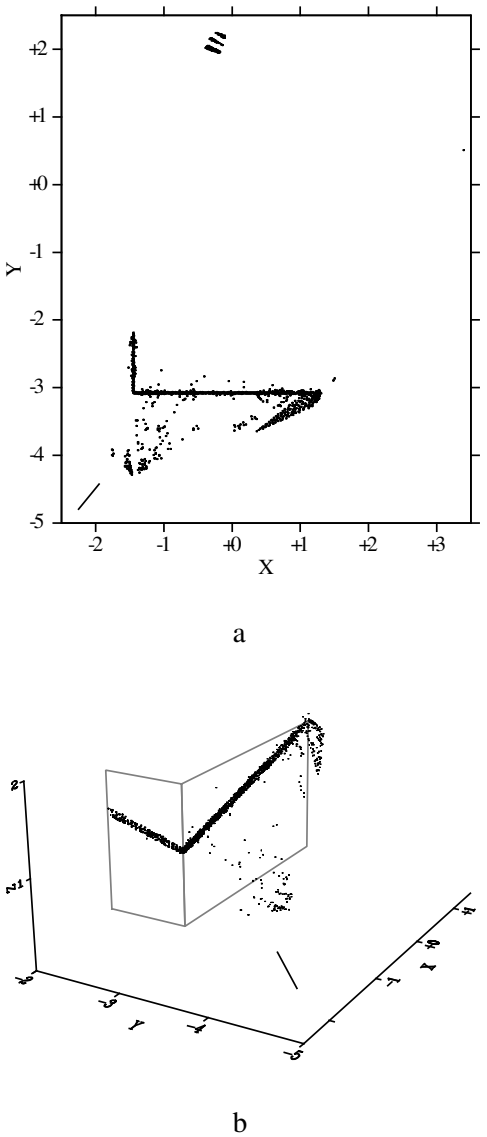

Figure 7: Reconstructed and actual points.

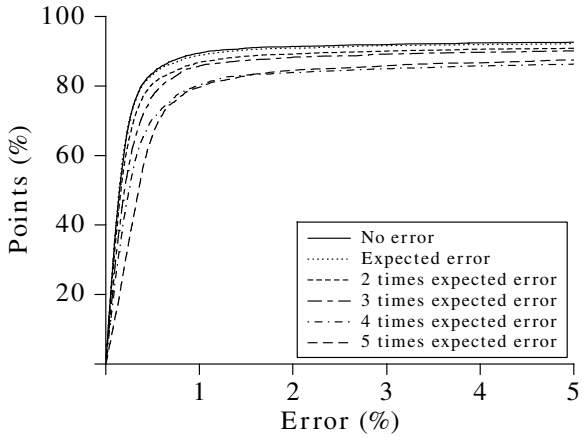

Figure 8: Number of points versus error.

Next, we considered outliers. Figure 8 shows the cumulative distribution of reconstruction error. Error is expressed as a percentage of the distance between the reconstructed $\mathrm{P_r}$ and actual $\mathrm{P_a}$ position divided by the depth of the reconstructed point ($\|\mathrm{P_r} - \mathrm{P_a}\| / \|\mathrm{P_r} - \mathrm{C_\star}\|$). The plotted curve indicates that the percentage[10] of reconstructed points with an error of less than 1% is 90% for the noise free case and 80% for noise levels of five times expected. Outliers also tend to have less support (fewer cameras contributing to the solution) and poor match quality (smaller values for $\nu(j,\alpha)$). Figure 9 shows the result of considering only points which have at least n cameras contributing to the solution. Similar results are obtained when points with small $\nu(j,\alpha)$ are removed.

Finally, we reconstructed another region about twice the size of the previous one which contained only building points. This time, we retained only points with more than 6 cameras contributing or with $\nu(j,\alpha) > -0.5$. Figure 10 shows the reconstructed points[11] rendered as oriented rectangular surface elements or *surfels* [Szeliski and Tonnesen, 1992]. We anticipate that the estimated orientation will prove very useful in fitting models to the reconstructed points or grouping them into surfaces.

[10] Sky points are omitted.

[11] Actually the data is downsampled by three in each direction for clarity.

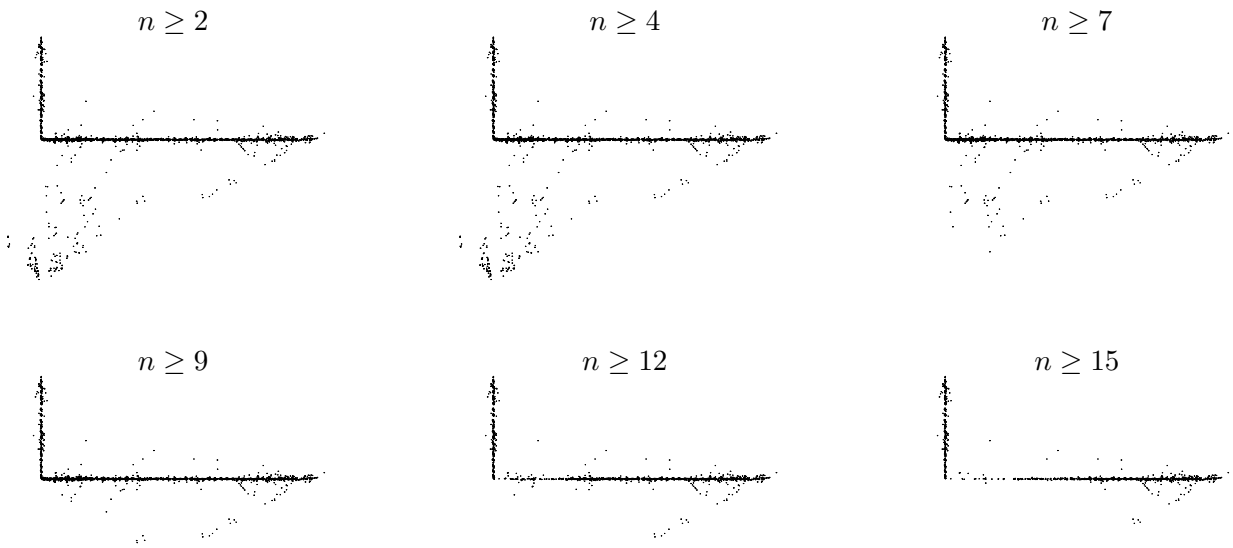

Figure 9: Outliers versus number of contributing cameras.

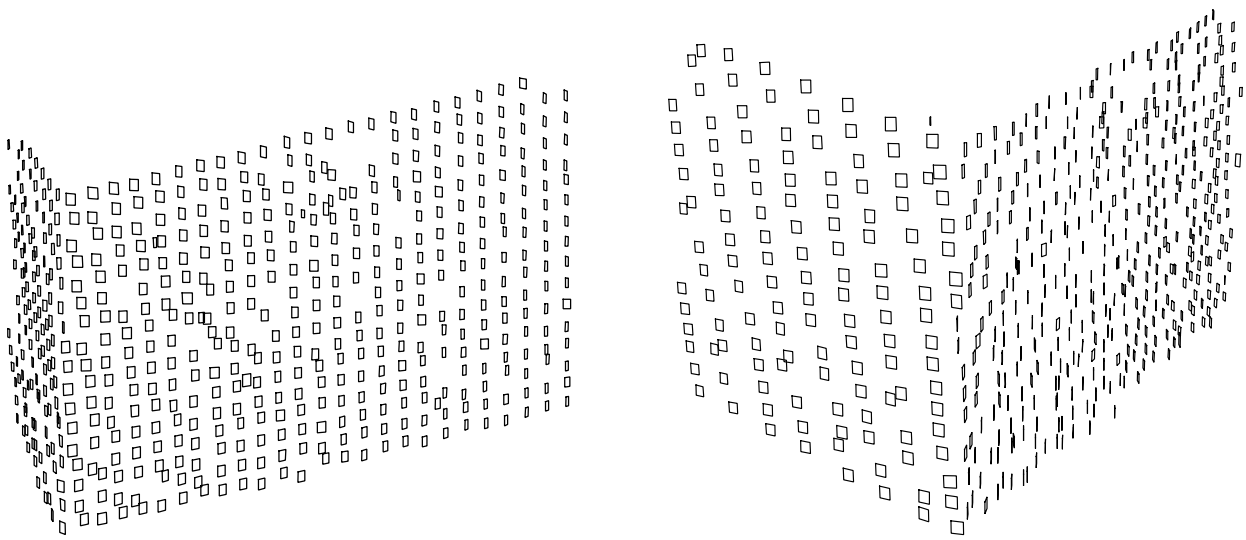

Figure 10: Two views of the reconstructed surfels.

4 Conclusions

This paper describes a method for generating dense depth maps directly from large sets of images taken from arbitrary poses. The algorithm presented is simple, accurate, and uses only local calculations. Our method builds, then analyzes, an epipolar image to accumulate evidence about the depth at each image pixel. This analysis produces an evidence versus depth and surface normal distribution that in many cases contains a clear and distinct global maximum. The location of this peak determines depth and orientation, and its shape can be used to estimate the error. The distribution can also be used to perform a maximum likelihood fit of models directly to the images. We anticipate that the ability to perform maximum likelihood estimation from purely local calculations will prove extremely useful in constructing three-dimensional models from large sets of images.

References

[Bolles *et al.*, 1987] Robert C. Bolles, H. Harlyn Baker, and David H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.

[Cox *et al.*, 1996] Ingemar J. Cox, Sunita L. Hingorani, Satish B. Rao, and Bruce M. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, May 1996.

[Duda and Hart, 1973] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, NY, 1973.

[Mellor *et al.*, 1996] J.P. Mellor, Seth Teller, and Tomás Lozano-Pérez. Dense depth maps from epipolar images. Technical Report AIM–1593, MIT, November 1996.

[Okutomi and Kanade, 1993] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, April 1993.

[Szeliski and Tonnesen, 1992] Richard Szeliski and David Tonnesen. Surface modeling with oriented particle systems. In *Computer Graphics (SIGGRAPH '92 Proceedings)*, volume 26, pages 185–194, July 1992.

[Tsai, 1983] Roger Y. Tsai. Multiframe image point matching and 3-D surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):159–174, March 1983.

[Yachida, 1986] M. Yachida. 3D data acquisition by multiple views. In O. D. Faugeras and G. Giralt, editors, *Robotics Research: the Third International Symposium*, pages 11–18. MIT Press, Cambridge, MA, 1986.