MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1593 November 1996

Dense Depth Maps from Epipolar Images

J.P. Mellor, Seth Teller and Tomás Lozano-Pérez

Abstract

Recovering three-dimensional information from two-dimensional images is the fundamental goal of stereo techniques. The problem of recovering depth (three-dimensional information) from a set of images is essentially the correspondence problem: *Given a point in one image, find the corresponding point in each of the other images*. Finding potential correspondences usually involves matching some image property. If the images are from nearby positions, they will vary only slightly, simplifying the matching process.

Once a correspondence is known, solving for the depth is simply a matter of geometry. Real images are composed of noisy, discrete samples, therefore the calculated depth will contain error. This error is a function of the baseline or distance between the images. Longer baselines result in more precise depths. This leads to a conflict: short baselines simplify the matching process but produce imprecise results; long baselines produce precise results but complicate the matching process.

In this paper, we present a method for generating dense depth maps from large sets (1000's) of images taken from arbitrary positions. Long baseline images improve the accuracy. Short baseline images and the large number of images greatly simplifies the correspondence problem, removing nearly all ambiguity. The algorithm presented is completely local and for each pixel generates 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 extremely useful in constructing three dimensional models from large sets of images.

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 Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-91-J-4038. The authors were also supported by the Advanced Research Projects Agency of the Department of Defense under Rome Laboratory contract F3060-94-C-0204.

1 Introduction

Recovering three-dimensional information from two-dimensional images is the fundamental goal of stereo techniques. The problem of recovering the missing dimension, depth, from a set of images is essentially the correspondence problem: *Given a point in one image find the corresponding point in each of the other images.* Finding potential correspondences usually involves matching some image property in two or more images. If the images are from nearby positions, they will vary only slightly, simplifying the matching process.

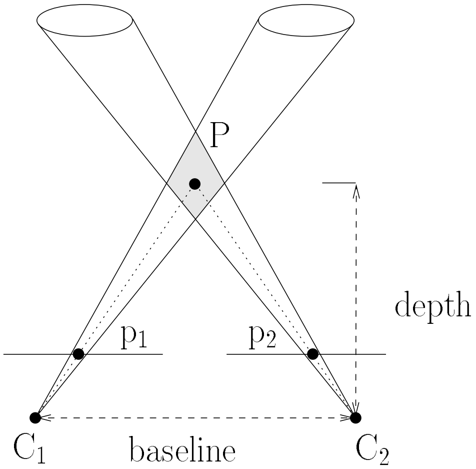

Figure 1: Stereo calculation.

Once a correspondence is known, solving for depth is simply a matter of geometry. Real images are noisy, and measurements taken from them are also noisy. Figure 1 shows how the depth of point P can be calculated given two images taken from known cameras C_1 and C_2 and corresponding points p_1 and p_2 within those images, which are projections of P. The location of p_1 in the image is uncertain, as a result P can lie anywhere within the left cone. A similar situation exists for p_2. If p_1 and p_2 are corresponding points, then P could lie anywhere in the shaded region. Clearly, for a given depth increasing the baseline between C_1 and C_2 will reduce the uncertainty in depth. This leads to a conflict: short baselines simplify the matching process, but produce uncertain results; long baselines produce precise results, but complicate the matching process.

One popular set of approaches for dealing with this problem are relaxation techniques[1] [6, 9]. These methods are generally used on a pair of images; start with an *educated guess* for the correspondences; then update them by propagating constraints. These techniques don't always converge and don't always recover the correct correspondences. Another approach is to use multiple images. Several researchers, such as Yachida [11], have proposed trinocular stereo algorithms. Others have also used special camera configurations to aid in the correspondence problem, [10, 1, 8]. Bolles, Baker and Marimont [1] proposed constructing an epipolar-plane image from a large number of images. In some cases, analyzing the epipolar-plane image is much simpler than analyzing the original set of images. The epipolar-plane image, however, is only defined for a limited set of camera positions. Tsai [10] and Okutomi and Kanade [8] defined a cost function which was applied directly to a set of images. The extremum of this cost function was then taken as the correct correspondence. Occlusion is assumed to be negligible. In fact, Okutomi and Kanade state that they "invariably obtained better results by using relatively short baselines." This is likely the result of using a spatial matching metric (a correlation window) and ignoring perspective distortion. Both methods used small sets of images, typically about ten. They also limited camera positions to special configurations. Tsai used a localized planar configuration with parallel optic axes; and Okutomi and Kanade used short linear configurations. Cox *et al* [2] proposed a maximum-likelihood framework for stereo pairs, which they have extended to multiple images. This work attempts to explicitly model occlusions, although, in a somewhat ad hoc manner. It uses a few global constraints and small sets of images.

The work presented here also uses multiple images and draws its major inspiration from Bolles, Baker and Marimont [1]. We define a construct called an *epipolar image* and use it to analyze evidence about depth. Like Tsai [10] and Okutomi and Kanade [8] we define a cost function that is applied across multiple images, and like Cox [2] we model the occlusion process. There are several important differences,

[1] For a more complete and detailed analysis of this and other techniques see [5, 7, 4].

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 [3].

2 Our Approach

In this section, we review epipolar geometry and epipolar-plane images, then define a new construct called an epipolar image. We also discuss the construction and analysis of epipolar images. Stereo techniques typically assume that relative camera positions and internal camera calibrations are known. This is sufficient to recover the structure of a scene, but without additional information the location of the scene cannot be determined. We assume that camera positions are known in a global coordinate system such as might be obtained from GPS (Global Positioning System). Although relative positions are sufficient for the discussion in this section, global positions allow us to perform reconstruction incrementally using disjoint scenes. We also assume known internal camera calibrations. The notation we use is defined in Table 1.

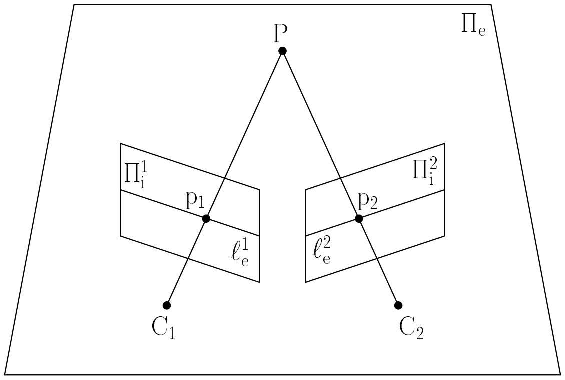

Figure 2: Epipolar geometry.

P_j	3D world point.
C_i	Center of projection for the i^{th} camera
$\Pi^i_{\rm i}$	Image plane.
p^j_i	Image point. Projection of P_j onto $\Pi^i_{\rm i}$.
$\Pi^k_{\rm e}$	Epipolar plane.
$\ell^i_{{\rm e},k}$	Epipolar line. Projection of $\Pi^k_{\rm e}$ onto $\Pi^i_{\rm i}$.
$\mathcal{EP}_k$	Epipolar plane image. Constructed using $\Pi^k_{\rm e}$.
$\mathrm{p}_\star$	Base image point. Any point in any image.
$\mathrm{C}_\star$	Base camera center. Camera center associated with $\mathrm{p}_\star$.
$\Pi^\star_{\rm i}$	Base image. Contains $\mathrm{p}_\star$.
$\ell_\star$	Base line. 3D line passing through $\mathrm{p}_\star$ and $\mathrm{C}_\star$.
$\mathcal{E}_k$	Epipolar image. Constructed using $\mathrm{p}_\star$. k indexes all possible $\mathrm{p}_\star$'s.
$\mathcal{F}(x)$	Function of the image at point x (e.g. image intensities, correlation window, features).
$\mathcal{X}(x_1, x_2)$	Matching function. Measures match between x_1 and x_2 (large value better match).
$\nu(j, \alpha)$	Match quality. Analyze $\mathcal{E}$.
$\{E \mid C\}$	Set of all E's such that C is true.
$\widehat{P_1 P_2}$	Unit vector in the direction from P_1 to P_2.
$\mathrm{d}(\mathrm{p}^j_i)$	Depth of image point p^j_i. If low confidence or unknown, then ∞.
M_l	Modeled object. Object whose position and geometry have already been reconstructed.

Table 1: Notation used in this paper.

2.1 Epipolar Geometry

Epipolar geometry provides a powerful stereo constraint. Given two cameras with known centers C_1 and C_2 and a point P in the world, the epipolar plane Π_e is defined as shown in Figure 2. P projects to p_1 and p_2 on image planes Π_i^1 and Π_i^2 respectively. The projection of Π_e onto Π_i^1 and Π_i^2 produces epipolar lines ℓ_e^1 and ℓ_e^2. This is the essence of the epipolar constraint. Given any point p on epipolar line ℓ_e^1 in image Π_i^1, if the corresponding point is visible in image Π_i^2, then it must lie on the corresponding epipolar line ℓ_e^2.

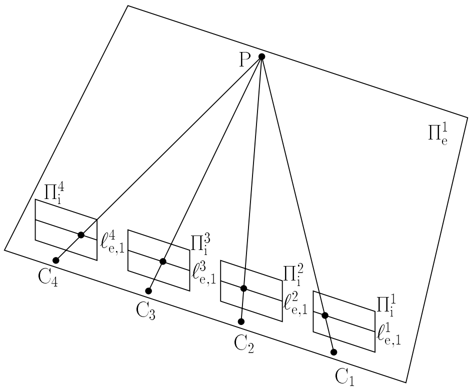

Figure 3: Epipolar-plane image geometry.

2.2 Epipolar-Plane Images

Bolles, Baker and Marimont [1] used the epipolar constraint to construct a special image which they called an epipolar-plane image. As noted earlier, an epipolar line ℓ_e^i contains all of the information about the epipolar plane Π_e that is present in the i^th image Π_i^i. An epipolar-plane image is built using all of the epipolar lines $\left\{\ell_{\mathrm{e},k}^i\right\}$ from a set of images $\{\Pi_\mathrm{i}^i\}$ which correspond to a particular epipolar plane Π_e^k (Figure 3). Since all of the lines $\left\{\ell_{\mathrm{e},k}^i\right\}$ in an epipolar-plane image $\mathcal{EP}_k$ are projections of the same epipolar plane Π_e^k, for any given point p in $\mathcal{EP}_k$, if the corresponding point in any other image Π_i^i is visible, then it will also be included in $\mathcal{EP}_k$. Bolles, Baker and Marimont exploited this property to solve the correspondence problem for several special cases of camera motion. For example, with images taken at equally spaced points along a linear path perpendicular to the optic axes, corresponding points form lines in the epipolar-plane image; therefore finding correspondences reduces to finding lines in the epipolar-plane image.

For a given epipolar plane Π_e^k, only those images whose camera centers lie on Π_e^k $\left(\left\{\mathrm{C}_i \mid \mathrm{C}_i\Pi_\mathrm{e}^k = 0\right\}\right)$ can be included in epipolar-plane image $\mathcal{EP}_k$. For example, using a set of images whose camera centers are coplanar, an epipolar-plane image can only be constructed for the epipolar plane containing the camera centers. In other words, only a single epipolar line from each image can be analyzed using an epipolar-plane image. In order to analyze all of the points in a set of images using epipolar-plane images, all of the camera centers must be collinear. This can be serious limitation.

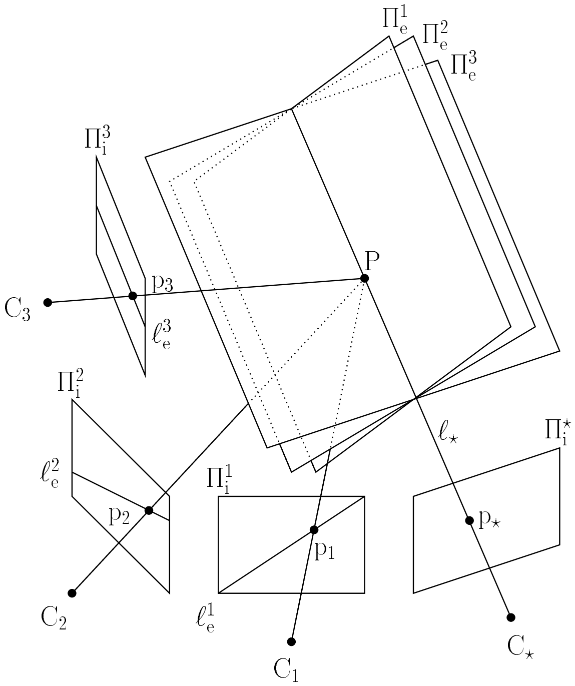

Figure 4: Epipolar image geometry.

2.3 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, 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 $\mathrm{C}_\star$ and one of the pixels $\mathrm{p}_\star$ in that image $\Pi_\mathrm{i}^\star$ (Figure 4). 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}. An epipolar-plane image is composed of projections of a plane; an epipolar image is composed of projections of a line. The cost of guaranteeing an epipolar image can be constructed for every pixel is that correspondence information is accumulated for only one point $\mathrm{p}_\star$, instead of an entire epipolar line.

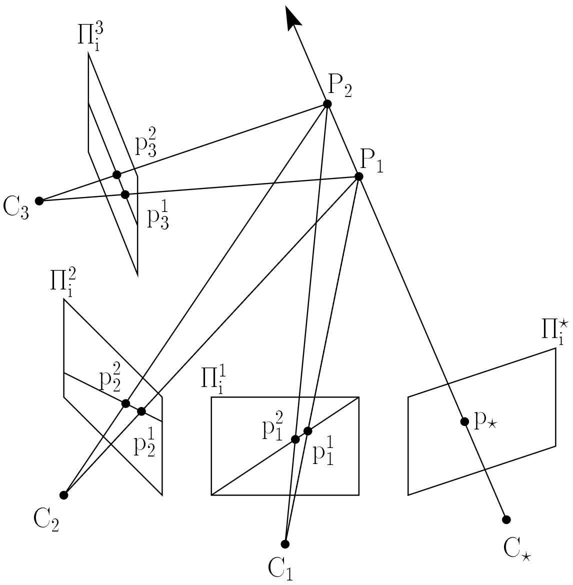

Figure 5: Set of points which form a possible correspondence.

To simplify the analysis of an epipolar image we can group points from the epipolar lines according to possible correspondences (Figure 5). P_1 projects to p_i^1 in Π_{i}^{i}; therefore $\{\mathrm{p}_i^1\}$ has all of the information contained in $\{\Pi_{\mathrm{i}}^{i}\}$ about P_1. There is also a distinct set of points $\{\mathrm{p}_i^2\}$ for P_2; therefore $\left\{\mathrm{p}_i^j \,|\, \text{for a given } j\right\}$ 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 $\{\Pi_{\mathrm{i}}^{i}\}$ and $\Pi_{\mathrm{i}}^{\star}$, then measurements taken at p_i^j should match those taken at $\mathrm{p}_\star$ (Figure 6a). 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 $\mathrm{p}_\star$ (Figures 6b and 6c). Epipolar images can be viewed as tools for accumulating evidence about possible correspondences of $\mathrm{p}_\star$. A simple function of j is used to build $\left\{\mathrm{P}_j \,|\, \forall i < j : \|\mathrm{P}_i - \mathrm{C}_\star\|^2 < \|\mathrm{P}_j - \mathrm{C}_\star\|^2\right\}$. In essence, $\{\mathrm{P}_j\}$ is a set of samples along $\ell_\star$ at in-

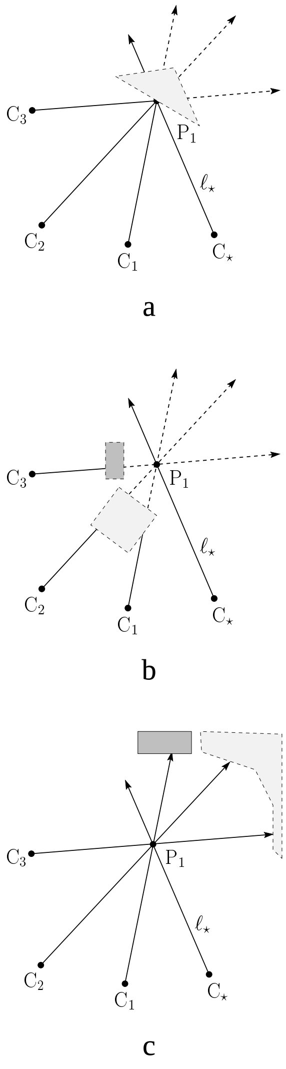

Figure 6: Occlusion effects.

creasing depths from the image plane.

2.4 Analyzing Epipolar Images

An epipolar image $\mathcal{E}$ is constructed by organizing

$$\left\{\mathcal{F}(\mathrm{p}_i^j) \mid \mathcal{F}() \text{ is a function of the image}\right\}$$

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[2] (Figure 7). 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 and j is its depth. Specifically:

$$\nu(j) = \sum_i \mathcal{X}(\mathcal{F}(\mathrm{p}_i^j), \mathcal{F}(\mathrm{p}_\star)), \qquad (1)$$

where $\mathcal{F}()$ is a function of the image and $\mathcal{X}()$ is a cost function which measures the difference between $\mathcal{F}(\mathrm{p}_i^j)$ and $\mathcal{F}(\mathrm{p}_\star)$. A simple case is,

$$\mathcal{F}(x) = \text{intensity values at } x$$

and

$$\mathcal{X}(x_1, x_2) = -|x_1 - x_2|.$$

Real cameras are finite, and p_i^j may not be contained in the image Π_i^i $\left(\mathrm{p}_i^j \notin \{\Pi_\mathrm{i}^i\}\right)$. Only terms for which $\mathrm{p}_i^j \in \{\Pi_\mathrm{i}^i\}$ should be included in (1). To correct for this, $\nu(j)$ is normalized, giving:

$$\nu(j) = \frac{\displaystyle\sum_{i \mid \mathrm{p}_i^j \in \{\Pi_\mathrm{i}^i\}} \mathcal{X}(\mathcal{F}(\mathrm{p}_i^j), \mathcal{F}(\mathrm{p}_\star))}{\displaystyle\sum_{i \mid \mathrm{p}_i^j \in \{\Pi_\mathrm{i}^i\}} 1}. \qquad (2)$$

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 $\left\{\mathrm{p}_i^j \mid \text{for a given } j\right\}$ increases, the likelihood increases that $\nu(j)$ will be large when P_j lies on a physical surface and small when it does not. Occlusions do not produce peaks at incorrect depths or false positives[3]. They can however, cause false negatives

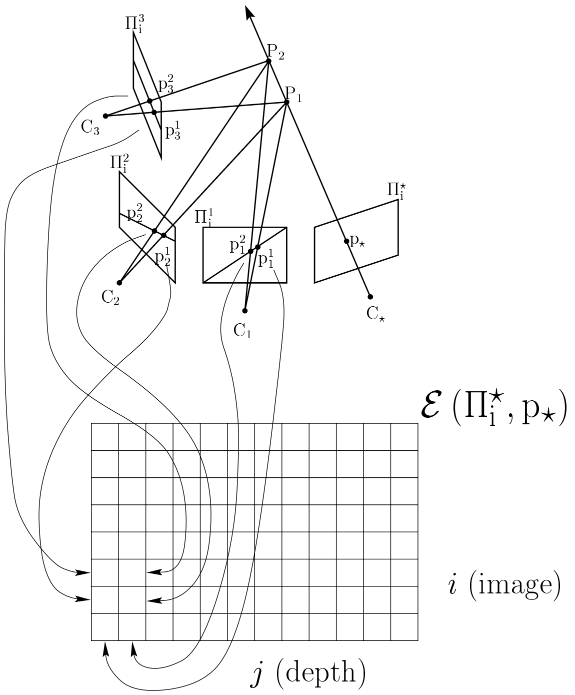

Figure 7: Constructing an epipolar image.

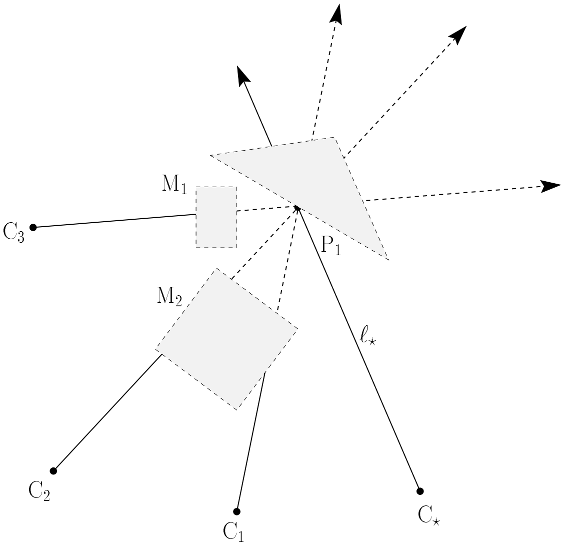

Figure 8: False negative caused by occlusion.

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

[3] Except possibly in adversarial settings.

or the absence of a peak at the correct depth (Figure 8). A false negative is essentially a lack of evidence about the correct depth. Occlusions can reduce the height of a peak, but a dearth of concurring images is required to eliminate the peak. Globally this produces holes in the data. While less then ideal, this is not a major issue and can be addressed in two ways: removing the contribution of occluded views, and adding unoccluded views by acquiring more images.

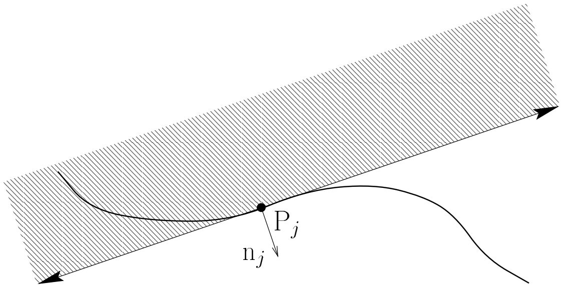

Figure 9: Exclusion region for P_j.

A large class of occluded views can be eliminated quite simply. Figure 9 shows a point P_j and its normal n_j. Images with camera centers in the hashed half space cannot possibly view P_j. n_j is not known a priori, but the fact that P_j is visible in $\Pi^\star_\mathrm{i}$ limits its possible values. This range of values can then be sampled and used to eliminate 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 $\widehat{\mathrm{C}_i\mathrm{P}}_j \cdot \alpha < 0$.

If the vicinity of $\{\Pi^i_\mathrm{i}\}$ is modeled (perhaps incompletely) by previous reconstructions, then this information can be used to improve the current reconstruction. Views for which the depth[4] $\mathrm{d}(\mathrm{p}^j_i)$ at p^j_i is less than the distance from Π^i_i to P_j can also be eliminated. For example, if M_1 and M_2 have already been reconstructed, then $i \in \{1,2,3\}$ can be eliminated from $\nu(j)$ (Figure 8). The updated function becomes:

$$\nu(j,\alpha) = \frac{\sum_{i\in\mathcal{S}} \mathcal{X}(\mathcal{F}(\mathrm{p}^j_i), \mathcal{F}(\mathrm{p}_\star))}{\sum_{i\in\mathcal{S}} 1} \qquad (3)$$

where

$$\mathcal{S} = \left\{ i \;\middle|\; \begin{array}{l} \mathrm{p}^j_i \in \{\Pi^i_\mathrm{i}\} \\ \widehat{\mathrm{C}_i\mathrm{P}}_j \cdot \alpha < 0 \\ \mathrm{d}(\mathrm{p}^j_i) \geq \|\mathrm{C}_i - \mathrm{P}_j\|^2 \end{array} \right\}.$$

Then, if sufficient evidence exists,

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

One way to eliminate occlusions such as those shown in Figure 8 is to process the set of epipolar images $\{\mathcal{E}_k\}$ in a best first fashion. This is essentially building a partial model and then using that model to help analyze the difficult spots. $\nu(j,\alpha)$ is calculated using purely local operations. Another approach is to incorporate global constraints.

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 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" (Figure 10). From this set of images, a $\Pi^\star_\mathrm{i}$ and $\mathrm{p}_\star$ were chosen and an epipolar image $\mathcal{E}$ constructed. $\mathcal{E}$ was then analyzed using two match functions:

$$\nu(j) = \frac{\displaystyle\sum_{i\,|\mathrm{p}^j_i \in \{\Pi^i_\mathrm{i}\}} \mathcal{X}(\mathcal{F}(\mathrm{p}^j_i), \mathcal{F}(\mathrm{p}_\star))}{\displaystyle\sum_{i\,|\mathrm{p}^j_i \in \{\Pi^i_\mathrm{i}\}} 1} \qquad (4)$$

and

$$\nu(j,\alpha) = \frac{\sum_{i\in\mathcal{S}} \left(\widehat{\mathrm{C}_i\mathrm{P}}_j \cdot \alpha\right) \mathcal{X}(\mathcal{F}(\mathrm{p}^j_i), \mathcal{F}(\mathrm{p}_\star))}{\sum_{i\in\mathcal{S}} \widehat{\mathrm{C}_i\mathrm{P}}_j \cdot \alpha} \qquad (5)$$

where

$$\mathcal{F}(x) = \mathrm{hsv}(x)^5 = [\mathrm{h}(x), \mathrm{s}(x), \mathrm{v}(x)]^\mathrm{T} \qquad (6)$$

[4] Distance from Π^i_i to the closest previously reconstructed object or point along the ray starting at C_i in the direction of p^j_i.

[5] hsv is the well known hue, saturation and value color model.

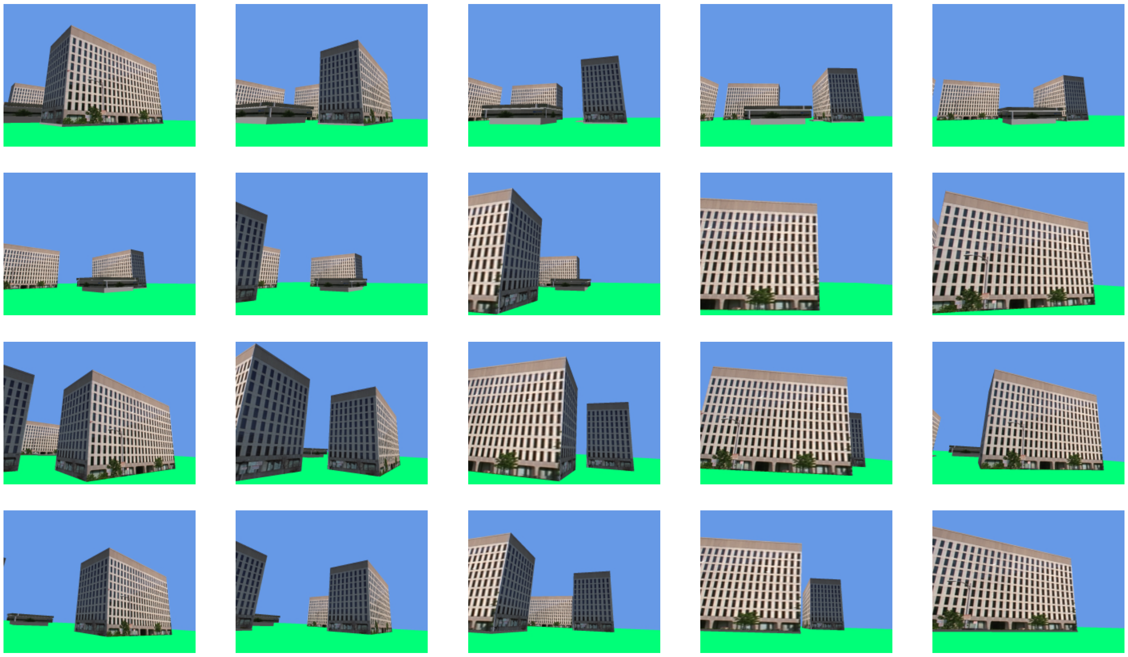

Figure 10: Examples of the rendered model.

$$\mathcal{X}([h_1,s_1,v_1]^{\mathrm{T}},[h_2,s_2,v_2]^{\mathrm{T}}) = -\left(\frac{s_1+s_2}{2}\right)(1-\cos(h_1-h_2)) - (2-s_1-s_2)\,|v_1-v_2|\,. \qquad (7)$$

Figures 11 and 12 show 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 $\nu(j)$ (4) and $\nu(j,\alpha)$ (5). 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 to increase visibility. To the right, $\nu(j)$ and $\nu(j,\alpha)$ are also shown as two-dimensional plots[6].

Figure 11a 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 side to side 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 $[-1446,-3078,1846]^{\mathrm{T}}$.

Figure 11b shows the epipolar image that results when a point just on the dark side of the front left edge of the building is chosen as $\mathrm{p}_{\star}$. Again both $\nu(j)$ and $\nu(j,\alpha)$ have a single peak that agrees well with the depth obtained using manual correspondence. This time, however, the peaks are asymmetric and have much broader tails. This is caused by the high contrast between the bright and dark faces of the building and the lack of contrast within the dark face. The peak in $\nu(j,\alpha)$ is slightly better than the one in $\nu(j)$.

Figure 11c shows the epipolar image that results when a point just on the bright side of the front left edge of the building is chosen as $\mathrm{p}_{\star}$.

[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.

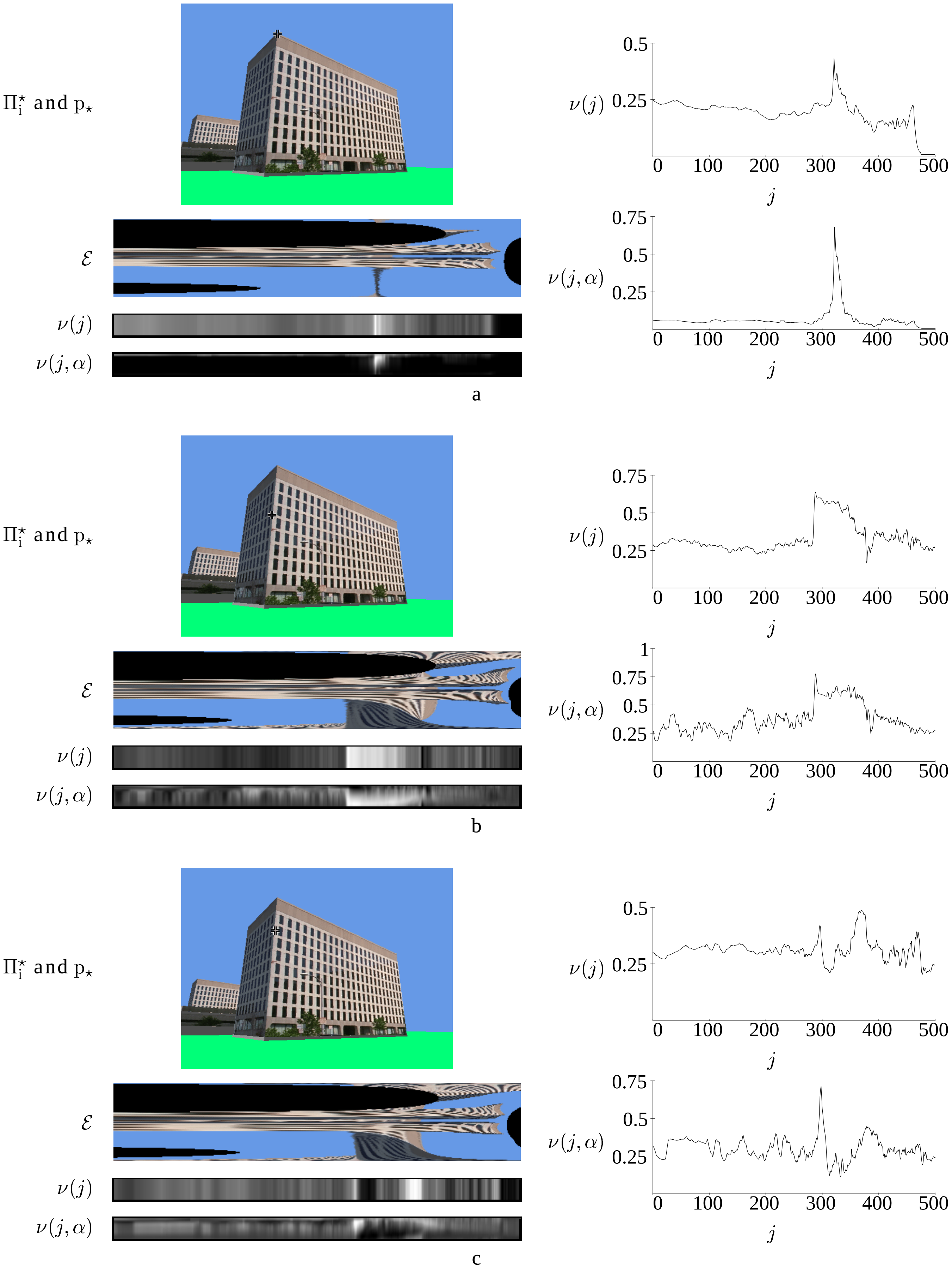

Figure 11: $\Pi_i^\star$, $p_\star$, $\mathcal{E}$, $\nu(j)$ and $\nu(j,\alpha)$.

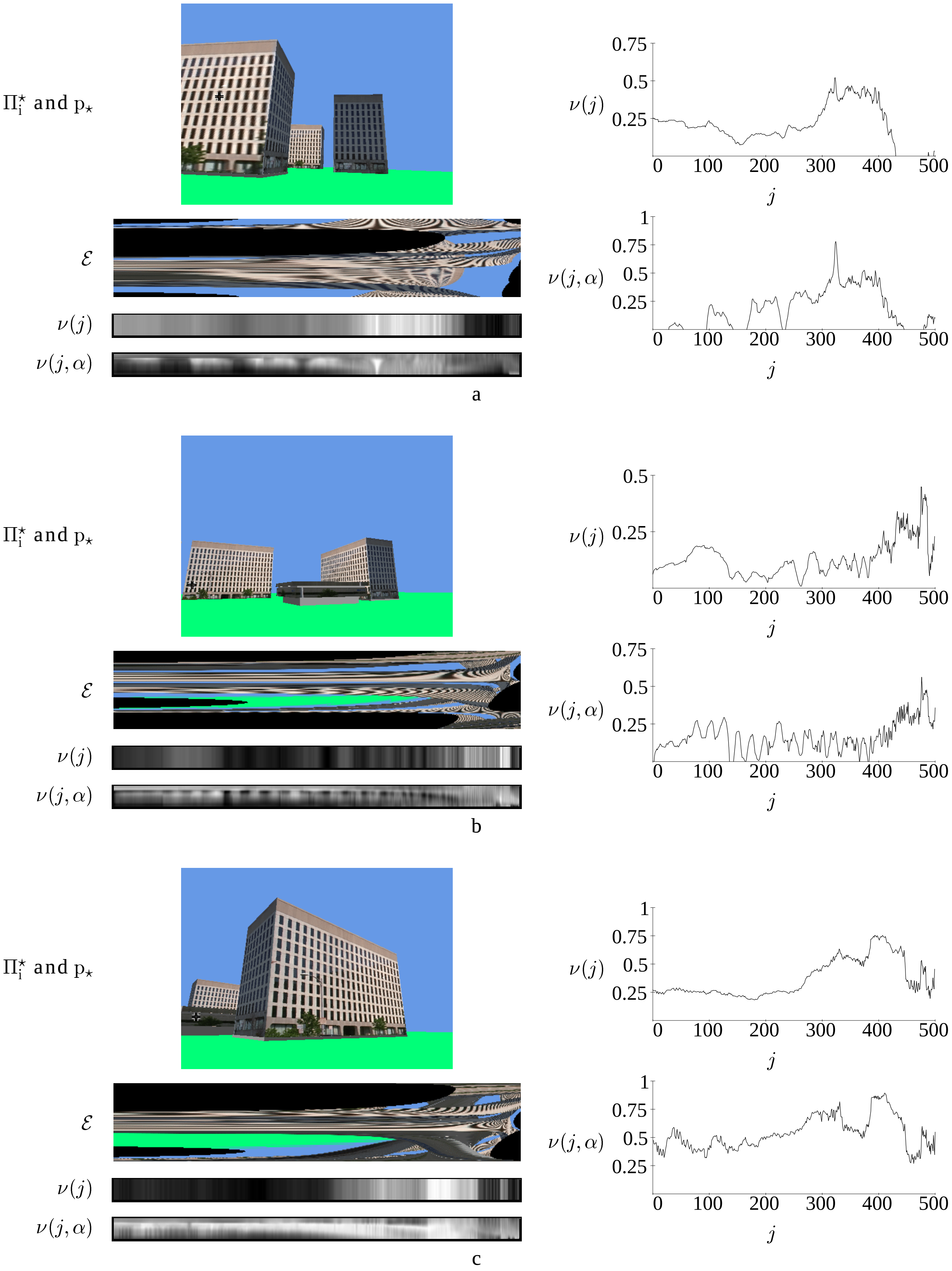

Figure 12: $\Pi_{\mathrm{i}}^{\star}$, $\mathrm{p}_{\star}$, $\mathcal{E}$, $\nu(j)$ and $\nu(j,\alpha)$.

This time $\nu(j)$ and $\nu(j,\alpha)$ are substantially different. $\nu(j)$ no longer has a single peak. The largest peak occurs at $j = 370$ and the next largest at $j = 297$. The manual measurement agrees with the peak at $j = 297$. The peak at $j = 370$ corresponds to the point where $\ell_\star$ exits the back side of the building. $\nu(j,\alpha)$, on the other hand, still has a single peak, clearly indicating the usefulness of estimating α.

In Figure 12a, $\mathrm{p}_\star$ is a point from the interior of a building face. 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 12b, $\mathrm{p}_\star$ is a point on a building face that is occluded (Figure 8) in a number of views. Both $\nu(j)$ and $\nu(j,\alpha)$ produce fairly good peaks that agree with manual measurements. In Figure 12c, $\mathrm{p}_\star$ is a point on a building face with very low contrast. In this case, neither $\nu(j)$ nor $\nu(j,\alpha)$ provide clear evidence about the correct depth. The actual depth occurs at $j = 386$. Both $\nu(j)$ and $\nu(j,\alpha)$ lack sharp peaks in large regions with little or no contrast or excessive occlusion. Choosing $\mathrm{p}_\star$ as a sky or ground pixel will produce a nearly constant $\nu(j)$ or $\nu(j,\alpha)$.

To further test our method, we reconstructed the depth of a region in one of the images (Figure 13). For each pixel inside the black rectangle the global maximum of $\nu(j,\alpha)$ was taken as the depth of that pixel. Figure 14a shows the depth for each of the 3000 pixels reconstructed plotted against the x image coordinate of the pixel. Slight thickening is caused by the fact that depth changes slightly with the y image coordinate. The cluster of points at the left end (near a depth of 7000) and at the right end correspond to sky points. The actual depth for each pixel was calculated from the CAD model. Figure 14b shows the actual depths (in grey) overlaid on top of the reconstructed values. Figure 15 shows the same data ploted in world coordinates[9]. The actual building faces are drawn in grey, and the camera position is marked by a grey line extending from the center of projection in the direction of the optic axis. The reconstruction shown (Figures 14 and 15) was performed purely locally at each pixel. Global constraints such as ordering or smoothness were not imposed, and no attempt was made to remove low confidence depths or otherwise post-process the global maximum of $\nu(j,\alpha)$.

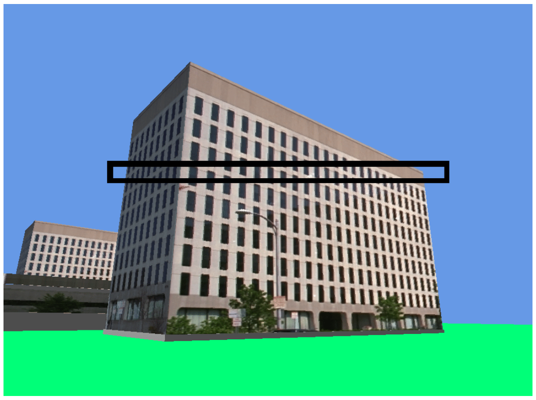

Figure 13: Reconstructed region.

4 Conclusions

This paper describes a method for generating dense depth maps directly from large sets of images taken from arbitrary positions. The algorithm presented uses only local calculations, is simple and accurate. 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 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 to the depth map. 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.

Acknowledgments

We wish to thank all those who helped construct the three-dimensional model that produced the synthetic imagery used in this paper. We also wish to thank George Chou for providing the rendered model and camera information.

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

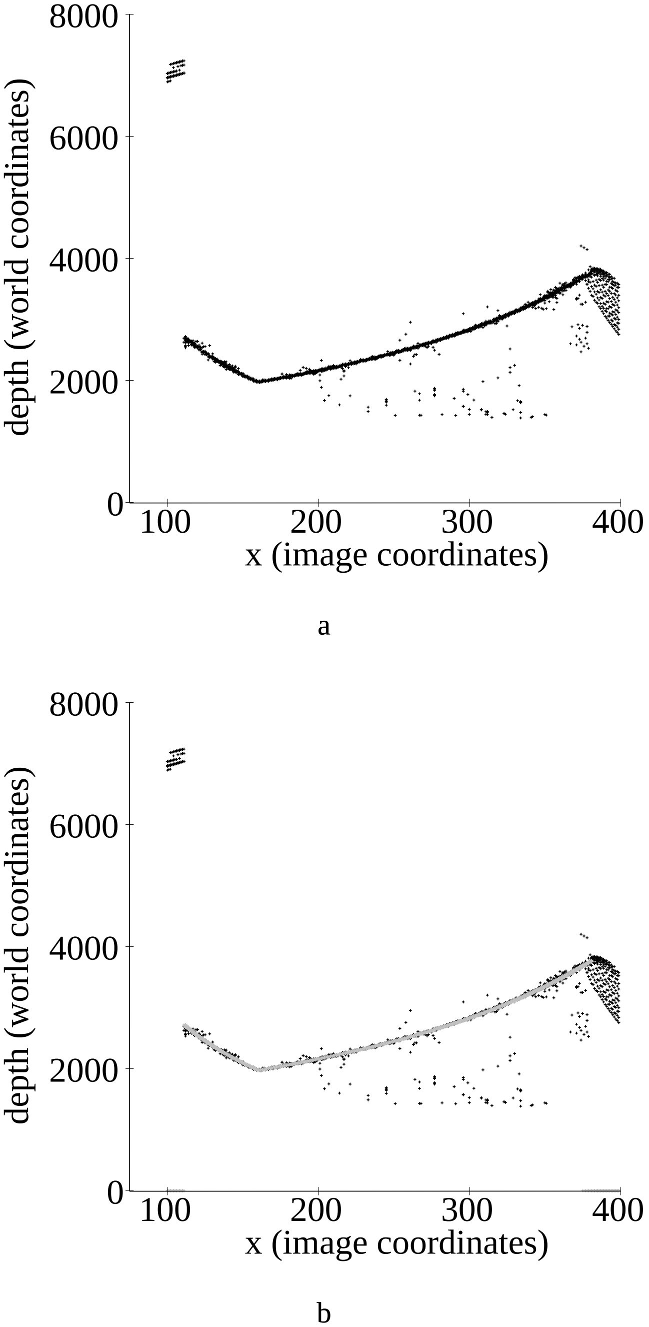

Figure 14: Reconstructed and actual depth maps.

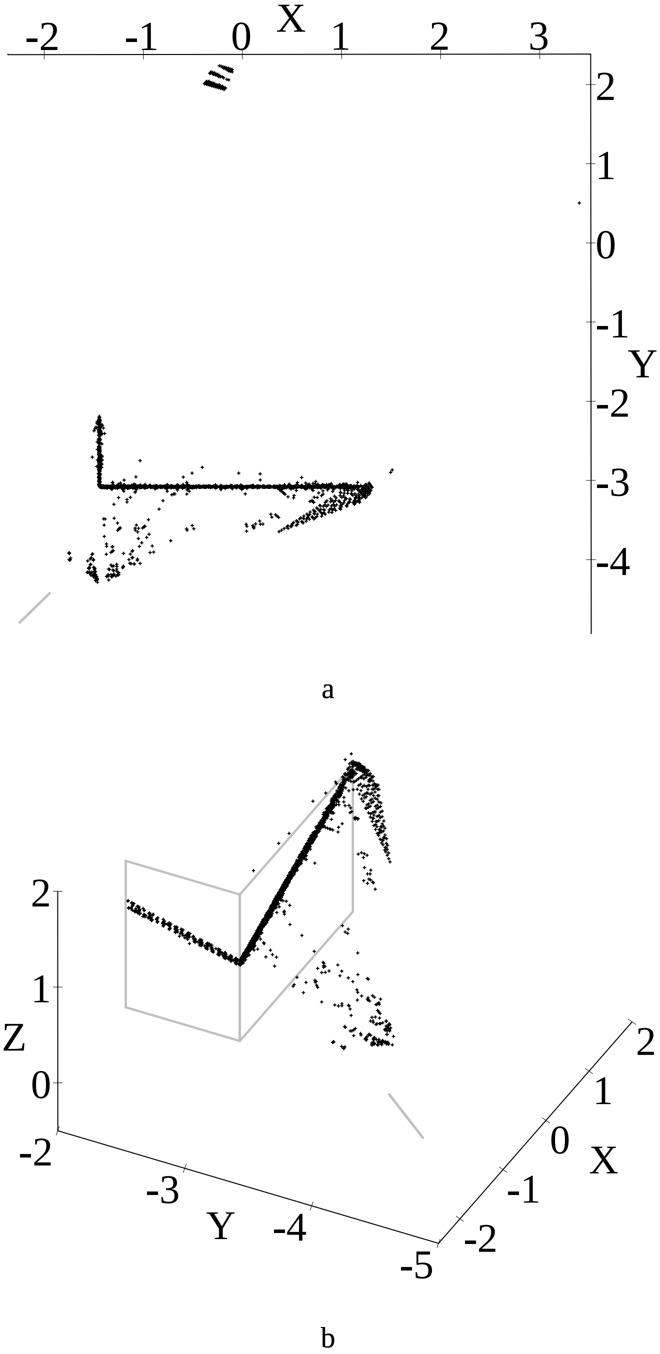

Figure 15: Reconstructed and actual world points.

References

[1] 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.

[2] 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.

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

[4] Olivier Faugeras. *Three-Dimensional Computer Vision*. MIT Press, Cambridge, MA, 1993.

[5] Berthold Klaus Paul Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.

[6] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London*, B(204):301–328, 1979.

[7] John E. W. Mayhew and John P. Frisby, editors. *3D Model Recognition from Stereoscopic Cues*. MIT Press, Cambridge, MA, 1991.

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

[9] S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby. Pmf: A stereo correspondence algorithm using a disparity gradient constraint. *Perception*, 14:449–470, 1985.

[10] 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.

[11] 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.