Dense Surface Patches from Thousands of Pose 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/city/city.html

Abstract

We are interested in recovering dense surface patches from large sets (1000's) of pose images (images annotated with absolute position and orientation) acquired from arbitrary locations within the scene. Previously we reported progress towards this goal using synthetic imagery, perfect calibration, controlled illumination conditions, and little or no occlusion. Here we extend our approach to use sets of images which 1) contain significant calibration errors, 2) are captured under widely varying illumination conditions, and 3) may include occlusion.

1 Introduction

Using multiple images to improving the results obtained by binocular stereo techniques is a natural extension. 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] or to provide an ordering constraint [Collins, 1996, Seitz and Dyer, 1997]. Some have proposed interactive methods [Debevec *et al.*, 1996] and probabilistic formulations [Cox *et al.*, 1996]. We have previously described [Mellor *et al.*, 1996] a method which uses large sets of pose images[1] to recover dense 3-D information. Our method differs from previous approaches in several important ways. It is valid for arbitrary camera positions and models some forms of occlusion. Our method is intended to recover dense surface patches or *surfels*[2] for built geometry (architectural facades) using thousands of images acquired from within the scene. In most cases, surfels can be recovered using purely local information, avoiding the computational costs of global constraints. In this paper we extend this approach to use sets of images which 1) contain significant calibration errors, 2) are captured under widely varying illumination conditions, and 3) may include occlusion.

2 Our Approach

A set of calibrated images were collected in and around a small office park. Nearly 4000

*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.

[1] Images annotated with absolute position and orientation.

[2] We estimate both 3-D position and normal for finite surface patches which are locally planar. We refer to these surface patches as surfels. It is a straight forward extension to consider other surface descriptions (e.g. surfels which locally are quadratic). Note our use of the term surfel differs from that in [Szeliski and Tonnesen, 1992].

images were collected from 81 node points [De Couto, 1998]. At each node, the camera was rotated about the focal point collecting images in a spherical mosaic. Most nodes are tiled with 47 images. The raw images are 1524×1012 pixels and cover a field of view of $44° \times 29°$. Next, the images are reduced to quarter size and their poses are refined [Coorg *et al.*, 1998].

2.1 Calibration Errors

After pose refinement, the calibration data is good, but not perfect. The pose estimates are within about 1° and 1 meter of the actual values. These errors produce an offset between corresponding points in different images. For example, a 1° pose error displaces a feature by over 8 pixels. Similar displacements are caused by position errors. Many n-camera stereo algorithms rely fundamentally upon reprojecting the images into a common frame. Random direction multi-pixel shifts tend to defeat these approaches. Figure 1 shows corresponding[3] subregions from several images reprojected onto the actual surface which gave rise to them. This collection is typical of our data set – it is neither the best nor the worst. The misalignment is in large part the result of errors in the pose estimates.

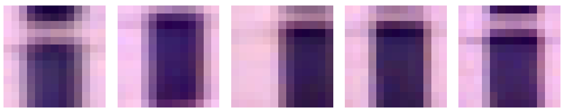

Figure 1: Reprojection onto actual surface.

Our method of recovering dense surfels from large sets of images involves finding peaks in the correlation of reprojected subregions. The naive approach for dealing with the misalignment would be to try all possible combinations of shifts. This requires testing an exponential number of combinations. Several improvements to the naive approach can be made. First, the set of all possible combinations contains numerous equivalent combinations. We avoid exploring most of the equivalent combinations by taking each image one at a time as a *key* image and finding the shift for each of the other images which aligns it best with the key image. Second, if the subregions are of sufficient size, gradient descent can be used for alignment. Finally, we can order the key images based on their *interestingness*. In practice, only the top few key images need to be explored and in any case only those with a least a minimum interest. The result is an algorithm which aligns corresponding textures and on the average is linear in the number of images and the maximum shift.

Once we have found a set of corresponding textures, we are still left with the question of where the textures came from. In general, there are many combinations of shifts which will align the textures, but the combinations are not equally likely. We assume that the pose errors are normally distributed and thus choose the maximum likelihood solution. The maximum likelihood shifts can also be used to further refine the pose estimates.

2.2 Varying Illumination

Our pose estimates are in an absolute coordinate frame which allows us to integrate images regardless of when or from what source they were collected. This greatly increases the quantity and quality of available data, but also complicates the analysis. The illumination conditions can vary widely (due to time of day, season, weather, etc.) from image to image. The textures shown in Figures 1 and 2 are all from the same reprojection plane and show the typical range of image variation[4].

We model this range of variation with a simple linear model. The radiance, R, arriving at an image sensor from a given region is simply the irradiance, I, at the region multiplied

[3] Determined using the estimated pose.

[4] The images in Figure 1 were all collected on bright, clear days. Those in Figure 2 were collected on overcast days and/or near sunset and have a distinct reddish tint.

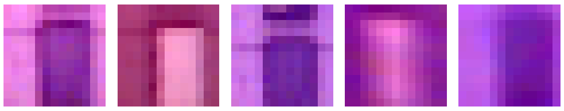

Figure 2: Differing illumination.

by its reflectance, r:

$$R = Ir \tag{1}$$

and the measured value is given by:

$$v = \mathrm{f}(R) + b, \tag{2}$$

Where f() is any invertible function and b is an offset. f() and b are functions of the image sensor. From Equations 1 and 2 it follows that if at some surface

$$I_1 r_1 = I_2 r_2 c \tag{3}$$

then

$$v_2 = \mathrm{f}(c) v_1 - \mathrm{f}(c) b_1 + b_2$$

or

$$v_2 = m v_1 + b. \tag{4}$$

Where x_i is the value of parameter x in illumination condition[5] n and c is an arbitrary constant. Equation 4 applies to both specular and diffuse surfaces. The condition given by Equation 3 requires that the product of irradiance and reflectance vary similarly for changes in illumination conditions at each point x in the region $\mathcal{X}$:

$$\forall x \in \mathcal{X} : I_1 r_1 / I_2 r_2 = \mathrm{c}. \tag{5}$$

In practical terms, this means that the region as viewed in condition 1 and 2 cannot contain shadows and limits mixing specular and diffuse materials in the same region. We assume that Equation 5 is satisfied and solve for m and b using the pixel values from the two regions[6]. Each color channel is done independently yielding a set of parameters that will map pixel values in condition 1 to pixel values in condition 2. If Equation 5 is not satisfied, then there will not be a good mapping and the match will be rejected.

[5] r is actually a function of viewing direction which we are lumping into illumination condition.

[6] For regions larger than 1×1 pixel Equation 4 is overdetermined; we find the least-squares solution.

2.3 Alignment and Color Correction

At each step of the alignment procedure described in Section 2.1 the color correction described in Section 2.2 is used. Figure 3 shows the results of applying these techniques to the reprojected textures in Figures 1 and 2.

Figure 3: Aligned and color corrected regions from Figures 1 (top) and 2 (bottom).

2.4 Optimization

As the reprojection plane is moved away from the actual surface, reprojected textures from different images become displaced from each other. As discussed above in Section 2.1, once we find a set of corresponding textures we choose the maximum likelihood position as the 3-D location for the texture. This implies a straightforward method for converging on the correct 3-D location by simple optimization. If the initial reprojection plane is not correct, only textures from images clustered relatively close together will match. We update the reprojection plane using the maximum likelihood estimate and try again. This time, textures from images a little further apart can contribute. The process continues until the set of textures and the 3-D position converge.

In a similar manner, the reprojection plane normal can be optimized using gradient information from the textures. Figure 4 depicts the algorithm. Images which form the set of corresponding textures and the reprojection plane are shown in black. Other images which were considered and rejected, lines of sight through the center of the matched textures, and building outlines are shown in various greys. Figure 4a shows the initial set

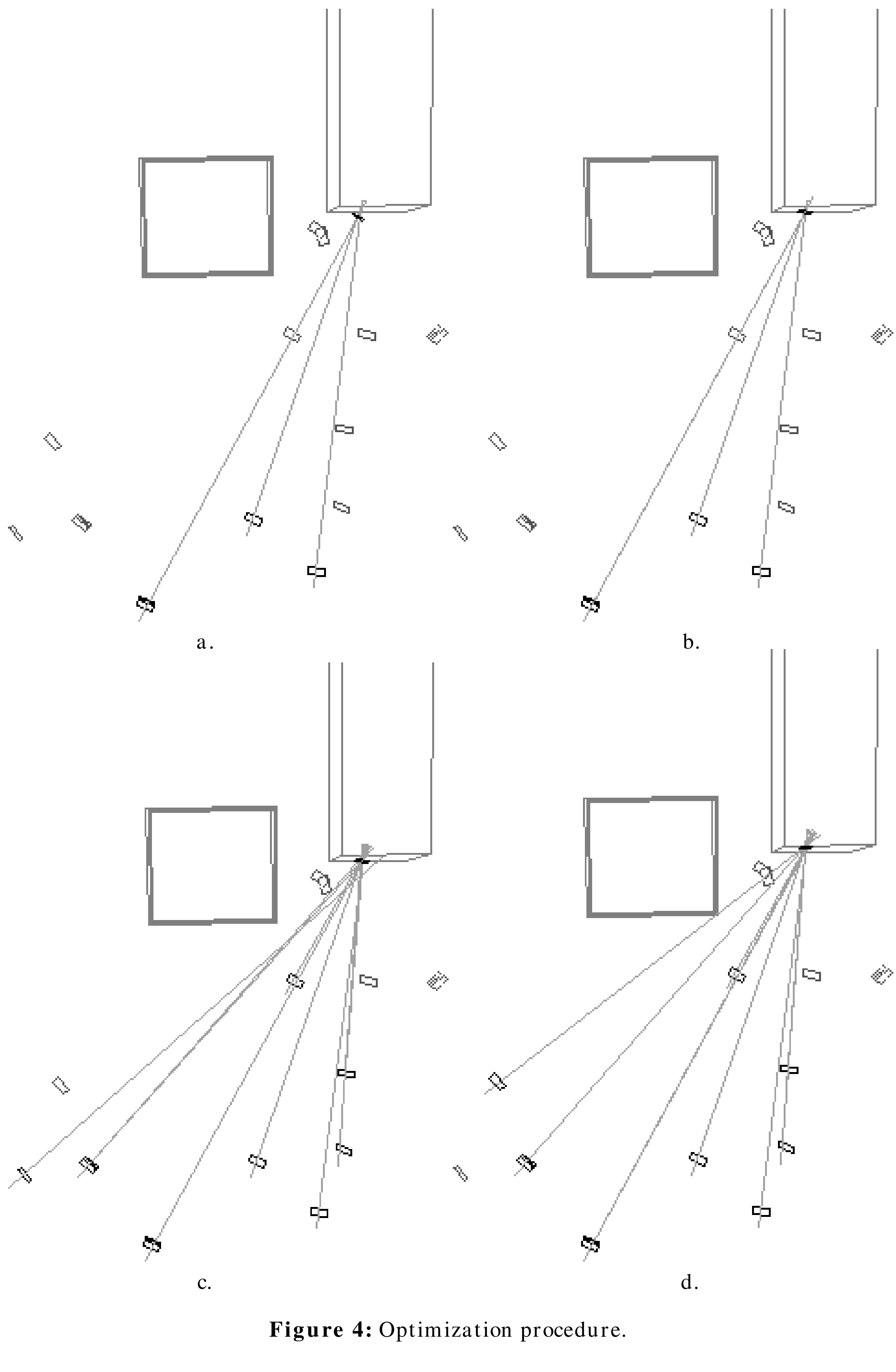

Figure 4: Optimization procedure.

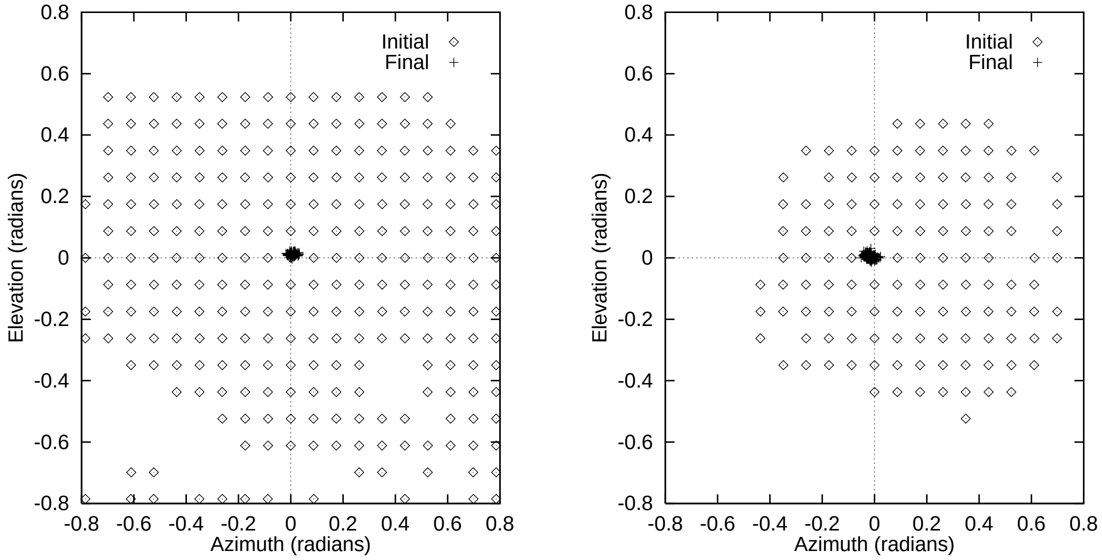

Figure 5: Initial and final (optimized) test points).

of corresponding textures. Figure 4b shows the initial set after updating the position and normal. Figure 4c shows a new set of corresponding textures using the updated postion and normal. Figure 4d shows the final set of corresponding textures at convergence.

To explore the convergence of our optimization method we tested initial reprojection planes every 20 units for positions ± 100 units, every $5°$ for elevations $\pm 45°$, and every $5°$ for azimuths $\pm 45°$ from the actual values. Figure 5 shows some of these results. The left plot is from 80 units in front of the actual surface and the right plot is from 80 units behind. The solution converges to correct values from a wide variety of starting normals. Initial and final values are plotted only for sets of corresponding textures which converged and matched sufficiently well.

2.5 Occlusion and Growing

Once a set of corresponding textures has been found and optimized, we can attempt to extend the solution. We simply reproject data for a larger area using the shifts and color corrections obtained using the initial set. For each point on the reprojection plane, we find the median pixel value[7]. If there are enough other pixels values close to the median value we retain the median value in our texture map. This approach allows us to use images with significant occlusion[8]. For example, Figure 6 contributes significantly to the median texture shown in 7. In fact it contributes almost everywhere the building isn't occluded by trees.

3 Conclusions

This paper describes extensions to an existing method for generating dense surface patches directly from thousands of pose images. These extensions allow the use of sets of images which 1) contain significant calibration errors, 2) are captured under widely varying illumination conditions, and 3) may include occlusion.

[7] By pixel value we mean RGB triplets. In general, pixel values can be n-vectors and contain data from additional spectrums.

[8] The use median values is similar to [Coorg, 1998].

Figure 6: Severely occluded image.

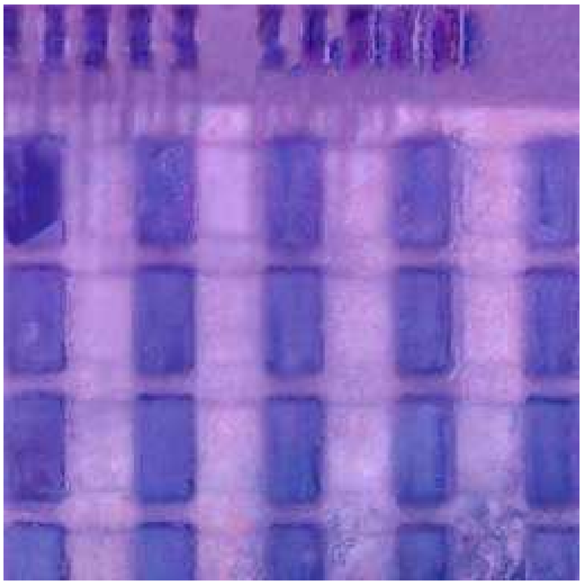

Figure 7: Median texture map.

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.

[Collins, 1996] Rober T. Collins. A space-sweep approach to true multi-image matching. In *Computer Vision and Pattern Recognition (CVPR '96 Proceedings)*, pages 358–363, June 1996. San Francisco, CA.

[Coorg *et al.*, 1998] Satyan Coorg, Neel Master, and Seth Teller. Acquisition of a large pose-mosaic dataset. In *Computer Vision and Pattern Recognition (CVPR '98 Proceedings)*, pages 872–878, June 1998.

[Coorg, 1998] Satyan Coorg. *Pose Imagery and Automated 3-D Modeling of Urban Environments.* PhD thesis, MIT, 1998.

[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.

[De Couto, 1998] Douglas S. J. De Couto. Instrumentation for rapidly acquiring pose-imagery. Master's thesis, MIT, 1998.

[Debevec *et al.*, 1996] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. Technical Report CSD–96–893, UCB, January 1996.

[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.

[Seitz and Dyer, 1997] Steven M. Seitz and Charles R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Computer Vision and Pattern Recognition (CVPR '97 Proceedings)*, pages 1067–1073, 1997.

[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.