Toward Urban Model Acquisition from Geo-Located Images

Seth Teller
Computer Graphics Group
MIT Lab for Computer Science[†]

Abstract

High-fidelity, textured geometric models are a fundamental starting point for computer graphics, simulation, visualization, design, and analysis. Existing tools for acquiring 3D models of large-scale (e.g., urban) geometry from imagery require significant manual input and suffer other, algorithmic scaling limitations. We are pursuing a research and engineering effort to develop a novel sensor, and associated geometric algorithms, to achieve fully automated reconstruction from close-range color images of textured geometric models representing built urban structures.

The sensor is a geo-located camera, which annotates each acquired digital image with metadata recording the date and time of image acquisition, and estimating the position and orientation of the acquiring camera in a global (geodetic) coordinate system. This metadata enables the formulation of reconstruction algorithms which scale well both with the number and spatial density of input images, and the complexity of the reconstructed model.

We describe our initial dataset of about four thousand geo-located images acquired through a prototype sensor, manual surveying, and semi-automated refinement of navigation information. We demonstrate, for a small office park on the MIT campus, the operation of fully automated algorithms for generating hemispherical image mosaics, for reconstructing vertical building facades, and for estimating high-resolution texture information for each facade. Finally, we describe the status of our efforts, and discuss several significant research and engineering challenges facing the project.

[†] *Cambridge MA 02139*

1. Introduction

Visual simulation systems require geometric models as input. Acquisition of complex 3D model geometry is a long-standing bottleneck in computer graphics and general simulation. We have chosen to address the problem of capturing models of urban scenes (i.e., textured building exteriors) from high-resolution color imagery. This is a natural problem domain in which to work, given the large number of applications engendered by rapid capture capability: emergency planning, urban planning, embedding of commercial databases, virtual sets, military operations, and the like. Also, the urban acquisition problem is rich in interesting research and engineering challenges.

We are addressing the acquisition problem with a research and engineering effort to develop a novel sensor and associated reconstruction algorithms [17]. We have developed a prototype mobile, geo-located digital camera to acquire digital images annotated with accurate estimates of acquisition time, camera position, and camera orientation. The availability of the sensor output, which we call "pose imagery," makes possible fundamentally novel and powerful approaches to the model reconstruction problem. The sensor and algorithms are described elsewhere [2, 6, 16, 9, 5, 7, 4].

1.1. Related Work

Existing computer vision algorithms [13, 11] usually have one or more significant limitations. Often, they make the assumption that input imagery is supplied already "controlled," or geo-registered. (Achieving geo-registration in existing photogrammetry systems is a manual labor-intensive task.) Typically, they are designed to operate on a small number of images, for example a stereo pair or triad, or a series of images acquired from a linear rail. Often the assumption is made that all images are pairwise correlated,

that is, that every pair of images observes some common portion of the real world. (This assumption clearly leads to algorithms with running times quadratic in the number of input images.) Alternatively, an image sequence (typically video) is processed, under the assumption that successive images are related. Both types of system usually require a human in the loop in order to initialize camera pose estimates, indicate or select features, or supply correspondences among features (e.g., [8]).

Such approaches are limited in at least three respects. First, they do not scale to large numbers of images or very complex output scenes, as human attention and handling are required for each image and/or each reconstructed feature. Second, they are not "algorithmic," as they require a human in the system to complete one or more tasks. Third, they produce model data in an arbitrary coordinate system used by the algorithm or chosen by the operator. In order to be useful outside of the modeling tool, a coordinate transformation must be effected, either manually or by the incorporation of photogrammetric fiducial points.

Researchers have demonstrated impressive reconstructions of roof and building structures from multiple controlled aerial images [3]. Another system has demonstrated high-quality reconstructions from imagery, but employs a human user to indicate block structures, feature correspondences, and unoccluded texture regions [8]. Recently, a reconstruction algorithm based on the evolution of level sets has appeared, with impressive results [12] for controlled imagery of textured shapes and faces. Another group has published a notion of "space carving" which produces a provably correct union of photo-consistent volumes from controlled imagery [15]. However, no fully automatic system for reconstruction of textured urban models from initially uncontrolled close-range imagery has yet been demonstrated.

1.2. Project Rationale

Our goal is the development of algorithms which extract from geo-located imagery a collection of textured geometric objects expressed in a global (Earth) coordinate system. We have demonstrated several algorithms which scale well with input and output complexity, using images of a small office park as a test dataset. These algorithms usefully exploit both the geo-locative metadata inherent in pose imagery, and the large number of images present in the dataset. The key strategies of this work are described below.

First, we augment existing digital cameras with **geo-locating sensors**, consisting of GPS (Global Positioning System) and inertial sensors, wheel encoders, a compass, and a Kalman filter which aggregates observations from disparate sensors. We are also experimenting with the incorporation of wide field-of-view video cameras in order to better estimate instantaneous translational and rotational velocities, and gather additional (low-resolution) observations.

Geo-location of each image allows us to insert incoming imagery directly into a spatial data structure indexed by position of the acquiring camera, and to query the data structure by region of interest. For example, in Figure 1, the query "report all cameras within 50 meters of, and containing within their field of view all or part of a specified region (the dashed rectangle)" might return the images shown in bold. Thus, for a given region of space, the number of images reported by the query (and thus subsequently processed) depends only on the size of the region of interest, and the density of images acquired in and around the region – but not on the total number of images in the dataset.

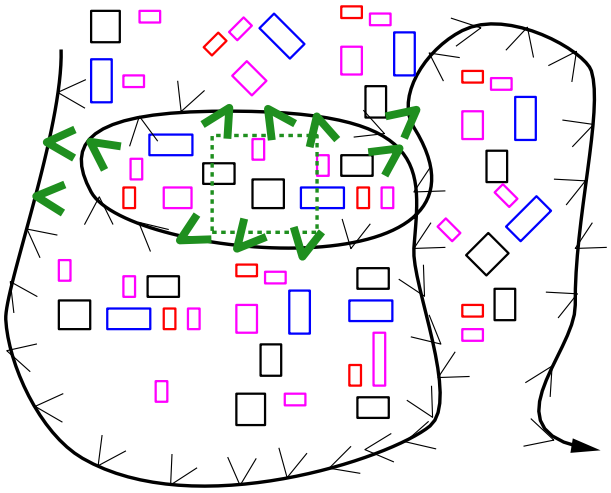

Figure 1. Pose metadata enables spatial indexing of large numbers of images, regardless of acquisition order.

This approach overcomes two scaling problems inherent in existing systems. Systems which use small numbers of images can employ brute-force algorithms to perform matching, but can achieve reconstruction data only over limited areas (for example, small portions of individual buildings). Moreover, by assuming all pairs of images are correlated, these algorithms have quadratic running times, so clearly cannot be applied to very large numbers of images. Systems which use large numbers of images (typically from a video sequence) assume that consecutive images are related, to perform reliable short-baseline tracking. However, these approaches have a different scaling problem: they must process *all* images before discovering *any*

correlated non-adjacent image pairs. To see this, consider an acquisition path that circles a region (Figure 1). In order to relate images far apart in the input sequence, all intermediate images must first be processed.

Our second strategy is to acquire **very large numbers (thousands) of high-resolution images**. The computer vision problem is generally underconstrained; that is, given a set of observations (images), there are many different three-dimensional structures and lighting conditions that could have given rise to it. Thus, acquiring many observations yields many constraints on the underlying model and illumination conditions which gave rise to those observations, improving confidence in the generated geometry and texture. Multiple observations help in a different fashion, as well. Significant structural elements (e.g. building corners, edges, or facades) will appear in many images. Our strategy is to aggregate many such observations so that real-world structures will have significant "signatures" in our aggregate. Preliminary results are quite encouraging. For example, from our first dataset of nearly 4,000 images [5] we can reliably detect the vertical facades of buildings in the Technology Square office park using a histogramming technique on edge orientations. Moreover, we can produce high-quality estimates of building texture for each facade, using a simple technique based on weighted median statistics. Both algorithms are described fully in [7, 4].

The third salient aspect, our use of **hemi-spherical images**, yields primarily engineering, rather than theoretical, advantages. We acquire imagery by rotating (panning and tilting) a camera into about 50 discrete orientations around a fixed optical center, or "node." Automated optimization techniques [5] aggregate the images into a single hemispherical mosaic image. The advantage is that, once the hemispherical mosaic is created, it can be treated as if it were captured by a physical camera with a hemispherical field of view.

The upshot of this strategy is that, when performing "bundle adjustment" (optimizing external calibration parameters) on the entire image set, the number of free variables to be optimized is reduced by a factor of fifty (to a single hemispherical image with 6 associated rigid DOFs, from fifty images, 6 DOFs each). Moreover, the effective resolution of the resulting mosaic image is nearly fifty times that of the underlying camera, clearly much higher than we could achieve with a single CCD array (even higher effective resolutions are achievable by further narrowing the underlying camera's field of view, at the cost of course of increased acquisition time). Both the reduced optimization load and increased image resolution are significant engineering advantages.

Fourth, the project is riding a collection of **technology trends** that make our methods timely. These trends are: the emergence of high-resolution electronic cameras (to acquire digital imagery); the availability of accurate GPS (positioning) and INS (orientation) sensors (to annotate imagery with geo-locative metadata); and the availability of ever more capacious storage devices and more powerful CPUs to store and process the data.

Finally, a significant portion of our work consists of **validation**, both of the accuracy of the geo-locative data produced by the sensor, and of the reconstruction data produced by our algorithms. To validate the sensor data, we acquire survey data through independent means consisting of manual surveying and the use of commercial photogrammetry systems. To validate the resulting reconstruction data we employ both survey data and traditional notions of image-space and world-space feature residuals.

2. Preliminary Results

This section describes the deployment of our prototype sensor in and around a small office park at MIT, and the subsequent processing of the acquired pose imagery.

2.1. Acquisition Platform

The pose camera [9] is a wheeled mobile platform with a high-resolution color digital camera mounted on a pan-tilt head, itself mounted on a vertically telescoping stalk. The platform also includes instrumentation to maintain estimates of global positioning (GPS), heading information (IMU), and dead-reckoning (mechanical wheel encoders). A Kalman filter maintains estimates of camera position and bearing in geodetic coordinates. Finally, an on-board power source and PC provide power and control to all of the devices, and a disk drive and digital tape drive store digital image and pose (position, heading, time) data.

2.2. First Dataset

We deployed an early prototype of the pose camera in and around Technology Square, an office park of four buildings located on the MIT campus. The prototype included a digital camera, mounted with fixed nodal point on a pan-tilt head, which was mounted on a moveable tripod. During acquisition, location and bearing information was derived from manually operated surveying instrumentation, and the introduction of one visible fiducial point for each node [5].

The pose camera was moved into eighty-one distinct locations. At each, a node was acquired by rotating the camera through a sequence of orientations. The resulting "tiling" amounted to a roughly hemispherical view of

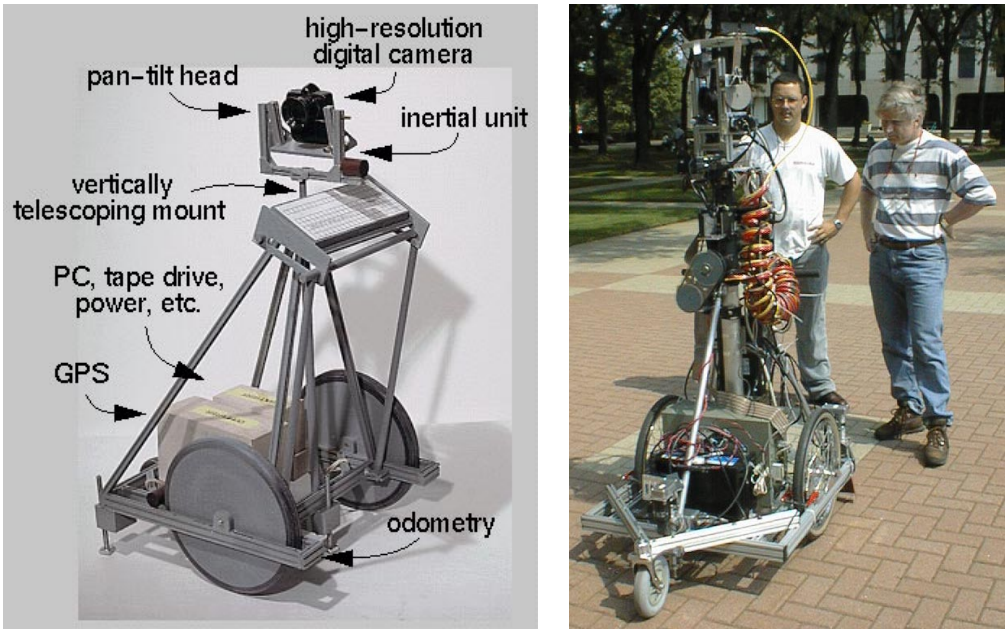

Figure 2. The prototype Argus platform (left: schematic; right: in the field).

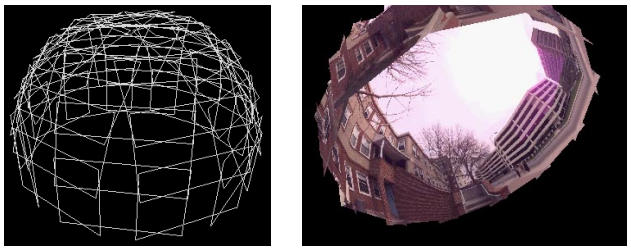

Figure 3. The hemispherical tiling (47 images) used for node acquisition.

the environment surrounding the camera, and comprised roughly fifty images (Figure 3). Digital imagery and metadata were transferred from the platform to our lab's storage facilities. At this stage, both position and rotation estimates were rough and in need of refinement; i.e., the imagery was only approximately controlled.

2.3. Mosaicing

The next step was to produce, from each set of images acquired at a node, a "spherical mosaic" which merged all of the images into a single, virtual image with a full hemispherical field of view. We do so with an iterative, dense correlation algorithm [5]. The algorithm's output is a quaternion (rotation) for each image, which relates that image to some reference direction for the node. Using these quaternions, and a interpolation scheme to blend pixels in overlap regions, virtual hemispherical images can be generated for each node location (Figure 4). At the right of the figure is another representation of the nodes; they have been unwrapped into cylinders, causing distortion of straight features (e.g., building edges).

2.4. Registration

Once each of the individual nodes has been processed into a mosaic, it remains to register the nodes with respect to each other; that is, to situate and orient each node in a common, global coordinate system. Our goal, as yet unrealized, is to eventually achieve a fully automated mechanism for exterior registration of the nodes. In the interim we employ a semi-automated method, which involves automatic identification of point features, and user selection and semi-automated correspondence detection [5]. Our tool registers all eighty-one nodes (comprising about 4,000 images) in about one hour of interaction time, or less than one second per image (Figure 5).

2.5. Reconstruction and Texture Estimation

Once we have a large collection of geo-registered digital images, a variety of 3D reconstruction algorithms can

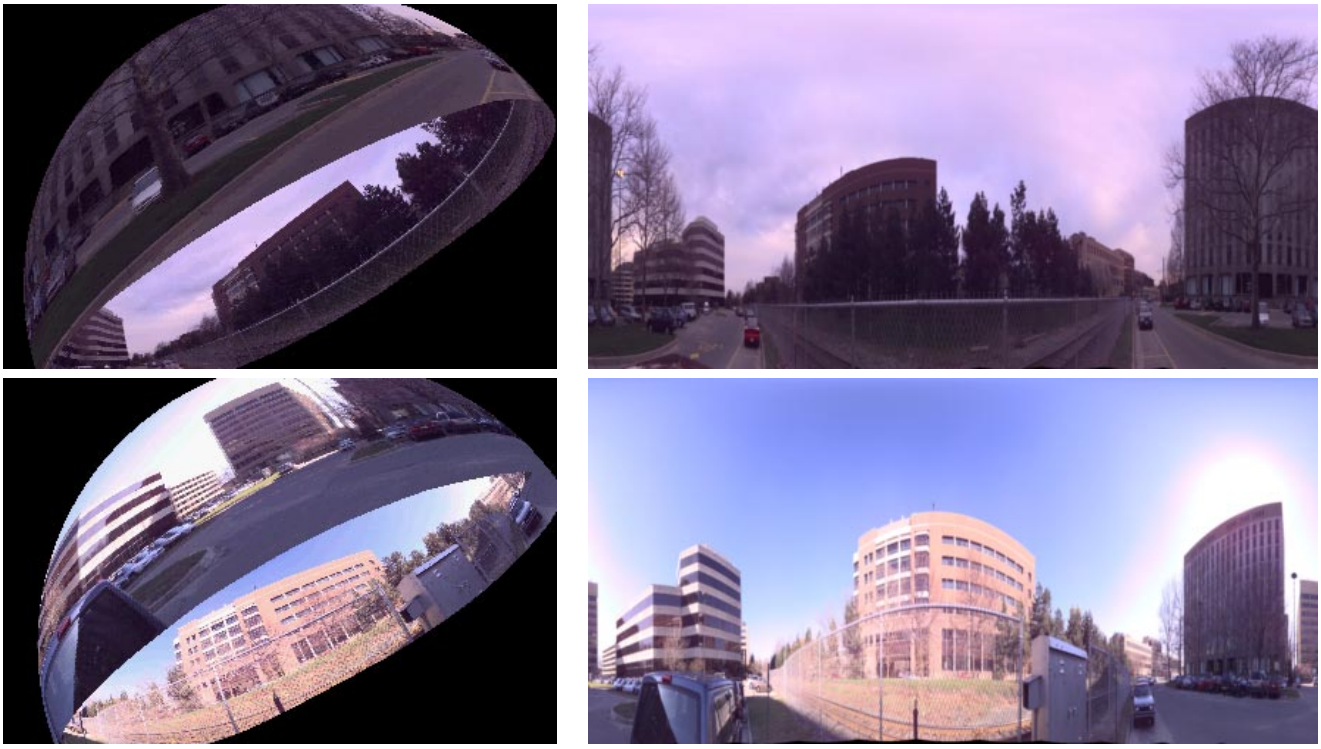

Figure 4. Two hemispherical mosaics (left), also displayed as cylinders (right).

be brought to bear. We are developing several complementary algorithms for inferring 3D structure from geo-located imagery. One algorithm matches dense regions of texture from many images to infer the existence of oriented surface elements [16]. Another algorithm generalizes temporal tracking methods to a spatial tracking method for sparse features such as corners, edges, and polygonal facades [2]. A third algorithm [1] establishes geometric variations within nearly-planar surfaces using plane plus parallax and level set evolution methods based on those described in [14, 8, 12].

A fourth algorithm is essentially mature [4]; it detects fragments of significant vertical facades via a Hough-like transformation [10] of horizontal scene edges. These fragments are linked into large vertical facades, which are then "extruded" downward to the ground terrain. Finally, median statistics are used to estimate for each facade a high-resolution texture map consistent with observation. A textured model generated by this algorithm is shown in Figure 6. The terrain is a triangulated height field, with heights derived from node survey information. It has been textured by the addition of a single geo-registered aerial image.

3. Challenges

This effort faces a number of engineering and research challenges. First, we must continue development of robust instrumentation for rapid acquisition of high-resolution digital imagery and navigation metadata. Interferometric differential GPS works well in open areas, but suffers from frequent dropouts in urban areas due to satellite occlusion and multipath interference, causing loss of phase lock. We are working to incorporate redundant navigation instrumentation (inertial sensors; wheel encoders; compass; etc.) with complementary performance characteristics to improve pose estimates.

Instrumentation improvements will produce more accurate, but not perfect, pose estimates for each acquired image. We anticipate that even with refined instrumentation, a node registration step will be required to produce controlled imagery of sufficient accuracy. Our semi-automatic methods are quite efficient; however, our goal is to achieve a fully automatic method.

From a systems point of view, the sheer data size of the input imagery and output models necessitates external-memory spatial indexing schemes to organize both the image data and generated models. We also employ multi-

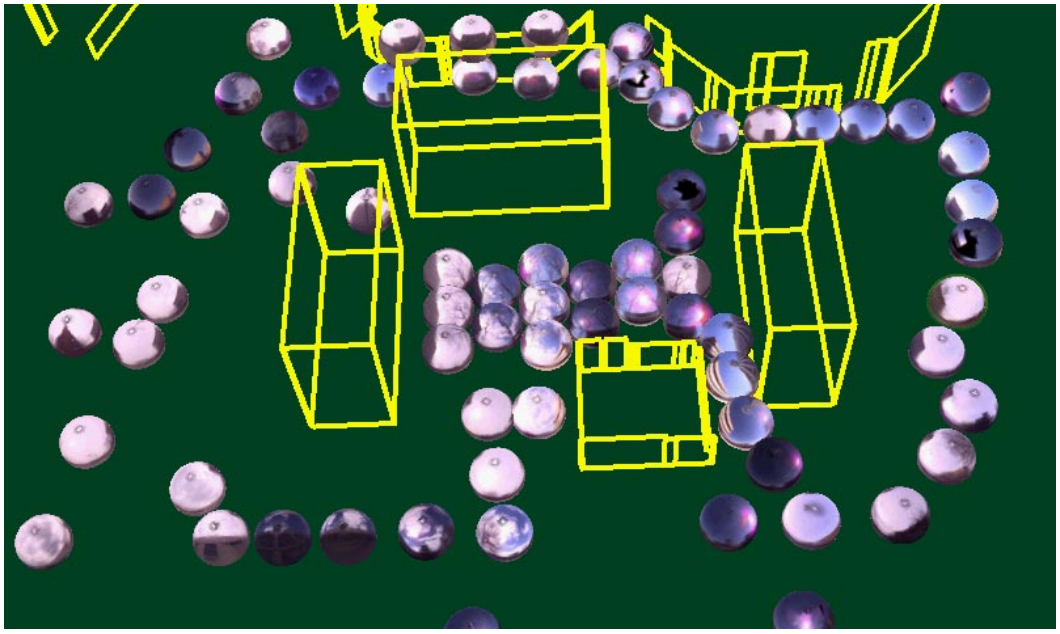

Figure 5. Initial pose-image dataset (about 4,000 images acquired from 81 node locations), geo-registered to a derived wireframe model of the Technology Square office park.

Figure 6. Result of reconstruction: a textured geometric model representing the office park.

scale techniques, for example performing initial pose refinement with low resolution imagery, then using the generated pose estimates as starting points for optimizations involving higher-resolution images.

Currently, our system implements techniques to reconstruct only vertical facades, which are "extruded" to ground and capped with rooftops using heuristics [4]. Clearly extensions will be required to capture, with reasonable fidelity, the rich variety of shapes present in an extended urban area.

The enormous variation in lighting conditions caused by changes in camera position, by occluding structures and foliage, by the passage of time, and by surface specularity, makes estimation of surface BRDFs difficult. We use weighted median statistics on multiple observations to eliminate most occluded pixels and estimate surface color [4]. This approach is naive in its assumption of constant, diffuse lighting and diffuse surfaces. Using the time estimates (and therefore associated sun positions and sky conditions) logged by our instrumentation, techniques such as those of [18] can be used to approximate reflectance properties of reconstructed surfaces.

4. Conclusion

We described the status of a project whose goal is fully automated capture of textured geometric models representing urban scenes. Imagery of a small office park demonstrated the prototype acquisition platform and reconstruction algorithms. All aspects of the system are fully automatic, save one: a semi-automated method for registering hemispherical images with respect to each other. The next goals for the project are to achieve full automation through improvement of both the sensor and the registration algorithms, and to scale the acquisition area up by two orders of magnitude, to a region about one kilometer square containing several hundred structures.

References

[1] E. Amram. A variational technique for three-dimensional reconstruction of local structure (in preparation). Master's thesis, Dept. of Electrical Engineering and Computer Science, MIT, 1998.

[2] G. Chou and S. Teller. Multi-image correspondence using geometric and structural constraints. In *DARPA Image Understanding Workshop*, May 1997.

[3] R. Collins, Y. Cheng, C. Jaynes, F. Stolle, X. Wang, A. Hanson, and E. Riseman. Site model acquisition and extension from aerial images. In *ICCV*, Cambridge, MA., 1995.

[4] S. Coorg. *Pose Imagery and Automated Three-Dimensional Modeling of Urban Environments (to appear)*. PhD thesis, MIT Ph.D. Thesis, 1998.

[5] S. Coorg, N. Master, and S. Teller. Acquisition of a large pose-mosaic dataset. In *CVPR '98*, pages 872–878, 1998.

[6] S. Coorg and S. Teller. Matching and pose refinement with camera pose estimates. *Proc. of DARPA IUW*, May 1997.

[7] S. Coorg and S. Teller. Automatic extraction of textured vertical facades from pose imagery. Technical Report TR-729, Laboratory for Computer Science, MIT, 1998.

[8] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *SIGGRAPH '96 Conference Proceedings*, pages 11–20, Aug. 1996.

[9] D. DeCouto. Instrumentation for rapidly acquiring pose imagery. Master's thesis, Dept. of Electrical Engineering and Computer Science, MIT, 1998.

[10] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[11] O. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, 1993.

[12] O. Faugeras and R. Keriven. Complete dense stereovision using level set methods. In *Proc. EECV*, pages 379–393, 1998.

[13] B. K. P. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.

[14] M. Irani, B. Rousso, and S. Peleg. Recovery of ego-motion using region alignment. *PAMI*, 19(3):268–272, March 1997.

[15] K. Kutulakos and S. Seitz. A theory of shape by space carving. Technical Report TR692, Computer Science Dept., U. Rochester, 1998.

[16] J. Mellor, S. Teller, and T. Lozano-Pérez. Dense depth maps for epipolar images. *Proc. of DARPA IUW*, May 1997.

[17] S. Teller. Automatic acquisition of hierarchical, textured 3D geometric models of urban environments: Project plan. In *Proceedings of the Image Understanding Workshop*, 1997.

[18] Y. Yu and J. Malik. Recovering photometric properties of architectural scenes from photographs. In *SIGGRAPH '98 Conference Proceedings*, pages 207–217, 1998.

(MIT publications available at `graphics.lcs.mit.edu`.)