PALM: Portable Sensor-Augmented Vision System for Large-Scene Modeling

Teck Khim Ng Takeo Kanade
The Robotics Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA

Abstract

We propose PALM – a Portable sensor-Augmented vision system for Large-scene Modeling. The system solves the problem of recovering large structures in arbitrary scenes from video streams taken by a sensor-augmented camera. Central to the solution method is the use of multiple constraints derived from GPS measurements, camera orientation sensor readings, and image features. The knowledge of camera orientation enhances computational efficiency by making a linear formulation of perspective ray constraints possible. The overall shape is constructed by merging smaller shape segments. Shape merging errors are minimized using the concept of shape hierarchy, which is realized through a "landmarking" technique. The features of the system include its use of a small number of images and feature points, its portability, and its low cost interface for synchronizing sensor measurements with the video stream. Example reconstructions of a football stadium and two large buildings are presented and these results are compared with the ground truth.

1. Background

The recovery of large scenes from video has an impact on applications such as architectural modeling and large scale virtual reality systems.

A large scene is by definition one that cannot be completely seen by a single camera view. Recovery of the complete scene requires the merging of smaller shape segments. The propagation and accumulation of merging errors is one of the most difficult problems of large scene reconstruction.

1.1. Structured Large Scenes

Merging errors can be reduced by using geometrical primitives if the scene is relatively structured. For example, very often the overall shape of the building can be constrained to be a rectangular block. This enforces a global shape constraint, which reduces the merging errors. *Facade*[2] is a successful system that adopts this approach. Geometrical primitives, such as rectangular blocks and prisms, are assigned manually to represent different parts of the structure.

Another way to reduce merging errors is to use a panorama created by image mosaicing. Shape merging errors are implicitly reduced when creating the mosaic in which a certain scene feature like a plane can be used to constrain the shape solution. This approach was adopted by Shum et al.[5]. They demonstrated accurate reconstruction of the interior structure of buildings.

1.2. Unstructured Large Scenes

Above mentioned systems, however, are not very effective in reconstructing large unstructured scenes, such as natural terrains. These scenes cannot be represented using simple geometrical primitives. The shape recovery needs to be done using structure-from-motion techniques.

Structure-from-motion for a large environment has two conflicting considerations. On one hand, it is desirable to make sure that each camera view sees a large portion of the structure so that the requirement for shape merging is minimal. On the other hand, keeping the large portion of the structure in view limits the amount of camera translations. Small camera translations, in turn, cause inaccuracies in structure-from-motion. It is also likely that the ratio of object depth to viewing distance will be large, making linear projection models invalid. Popular structure-from-motion methods like Factorization [6] and Extended Kalman Filtering [4, 1] will give inaccuracies in these cases.

In order to do structure-from-motion precisely, one is frequently forced to do a small portion of the structure at a time. The small shape segments need to be merged to form the complete big structure, and merging errors have to be dealt with.

0-7695-0062-5/99 $10.00 © 1999 IEEE

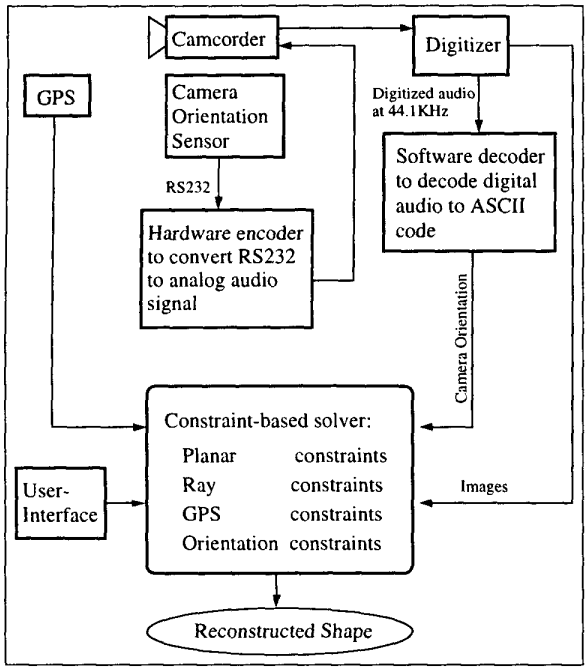


Figure 1. The PALM System

2. The PALM System

To solve the large scene reconstruction problem, we propose PALM – **P**ortable sensor **A**ugmented vision system for **L**arge-scene **M**odeling (Fig. 1). PALM does this by using multiple constraints derived from the use of auxiliary sensors (such as GPS and heading/tilt sensors) and image features. The use of auxiliary sensors helps to constrain the overall shape recovery process.

2.1. Portable Data Acquisition System

PALM has a portable data acquisition system (Fig. 2). The camcorder is mounted on top of a box that contains a low-cost camera orientation sensor and a hardware interface that we built to synchronize the sensor readings with the video stream. The internal parameters of the camcorder are calibrated using LaRose's method [3].

2.2. Feature Selection and Correspondence

In PALM, the user is required to specify planes, points and point correspondences using a graphical user-interface. However, in contrast to *Facade* [2] and Shum's system [5], which require interactive input from the user throughout the shape reconstruction process, PALM requires user input only at the beginning of the process.

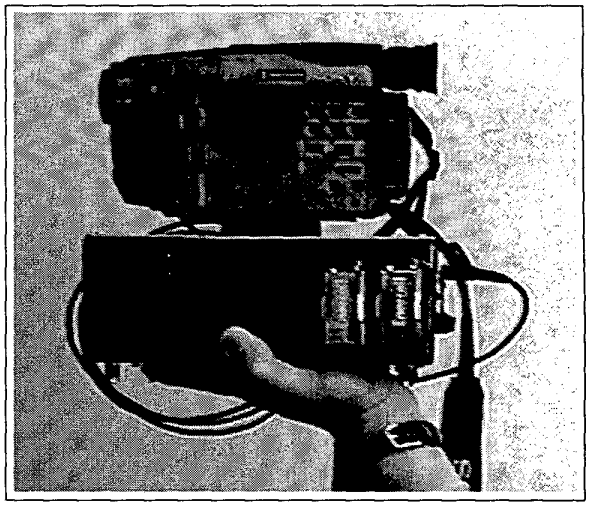


Figure 2. Data acquisition system of PALM

2.3. Linear solution of shape vector for complete structure

The use of the heading/tilt sensor achieves computational efficiency for the solution of the 3D shape. In particular, it allows the linear formulation of perspective ray constraints.

Ray constraints for all points in images are written using the familiar perspective projection equations (1) and (2):

$$\frac{l\,(\mathbf{p}_p - \mathbf{t}_f)\cdot \mathbf{i}_f}{(\mathbf{p}_p - \mathbf{t}_f)\cdot \mathbf{k}_f} = u_{fp} \quad (1)$$

$$\frac{l\,(\mathbf{p}_p - \mathbf{t}_f)\cdot \mathbf{j}_f}{(\mathbf{p}_p - \mathbf{t}_f)\cdot \mathbf{k}_f} = v_{fp} \quad (2)$$

where

l is the camera focal length,
$\mathbf{p}_p$ is the p^{th} shape point,
$\mathbf{t}_f$ is the camera translation for the f^{th} frame,
$\mathbf{i}_f$ is the camera horizontal axis direction for the f^{th} frame,
$\mathbf{j}_f$ is the camera vertical axis direction for the f^{th} frame,
$\mathbf{k}_f$ is the camera optical axis direction for the f^{th} frame,
u_{fp} is the horizontal image coord of p^{th} point in the f^{th} frame,
v_{fp} is the vertical image coord of p^{th} point in the f^{th} frame.

Equations (1) and (2) can be re-written respectively as:

$$(l\mathbf{i}_f - u_{fp}\mathbf{k}_f)\cdot \mathbf{p}_p = (l\mathbf{i}_f - u_{fp}\mathbf{k}_f)\cdot \mathbf{t}_f \quad (3)$$

$$(l\mathbf{j}_f - v_{fp}\mathbf{k}_f)\cdot \mathbf{p}_p = (l\mathbf{j}_f - v_{fp}\mathbf{k}_f)\cdot \mathbf{t}_f \quad (4)$$

Assume that frame f sees a total of c ($\geq$ 2) shape points. These c points can be concatenated into a $3c$ X 1 shape vector $\mathbf{x}_f = (\mathbf{p}_1^T\ \mathbf{p}_2^T\ \mathbf{p}_3^T\ \mathbf{p}_4^T\ \ldots\ \mathbf{p}_c^T)^T$.

Collecting all points in frame f, one can use (3) and (4) to construct the linear equation

$$B_f \mathbf{x}_f = A_f \mathbf{t}_f \quad (5)$$

where B_f is a $2c$ by $3c$ matrix and A_f is a $2c$ by 3 matrix.

The camera translation vector $\mathbf{t}_f$ can be written as

$$\mathbf{t}_f = (A_f^T A_f)^{-1} A_f^T B_f \mathbf{x}_f. \quad (6)$$

Vector $\mathbf{t}_f$ is therefore a linear combination of the elements of the shape vector $\mathbf{x}_f$. (5) is now written as

$$(B_f - A_f(A_f^T A_f)^{-1} A_f^T B_f)\mathbf{x}_f = 0 \quad (7)$$

Since PALM is equipped with a camera orientation sensor, $\mathbf{i}_f$, $\mathbf{j}_f$ and $\mathbf{k}_f$ can be derived from sensor readings. If point correspondences and camera focal length are known, the matrices A_f and B_f are completely specified. Therefore, by collecting all frames, all points and point correspondences, (7) forms a large linear system for the solution of the complete shape vector $\mathbf{x}$, where $\mathbf{x}$ is the column vector comprising all 3D shape points (i.e., formed by concatenating the non-repeating points of $\mathbf{x}_f$, for all f).

2.4. Planar scenes of known orientation

In some cases, additional constraints are available to be added to the linear system for the solution of the complete shape vector $\mathbf{x}$. For example, if a scene contains special features like planar configurations, planar constraints can be written.

Man-made objects like buildings usually have planar facades. In most cases, these surfaces are perpendicular to each other. For such scenes, it is easy for a user to specify the plane directions based on the building coordinate frame.

For example, for planes in one orientation, the plane normal can be assigned $\mathbf{n}_1 = [1\ 0\ 0]^T$. For other planes perpendicular to $[1\ 0\ 0]^T$, their normals can be $\mathbf{n}_2 = [0\ 0\ 1]^T$ or $[0\ 1\ 0]^T$. If, in addition, the camera orientation with respect to building frame is known, the 3D coordinates of these planar points can be recovered up to a scale ambiguity.

However, a scene may consist of different buildings. It is therefore necessary to use a common reference frame in order to refer to all planar directions. In PALM, the earth frame is chosen as the common reference frame (the heading/tilt sensor readings are measured with respect to this earth frame). The plane normal vectors $\mathbf{n}$ can be transformed to refer to this earth frame using

$$\mathbf{n}_i^E = (\mathbf{n}_i^T R_B^E)^T, \quad 1 \le i \le 2 \quad (8)$$

where R_B^E is the building orientation w.r.t. earth.

R_B^E can be obtained from the following equation:

$$R_B^E = R_S^E R_C^S (R_C^B)^{-1} \quad (9)$$

where R_S^E : sensor orientation w.r.t. earth,
R_C^S : camera orientation w.r.t. sensor,
R_C^B : camera orientation w.r.t. building.

R_S^E is the output of the orientation sensor, and R_C^S is obtained by calibrating the image-plane to the sensor.

R_C^B can be calculated if the building contains at least a pair of horizontal lines and a pair of vertical lines [5]. Note that R_C^B needs to be established this way for only one frame. This is because once R_B^E is determined, R_C^B for the rest of the frames for this building can be estimated using

$$R_C^B = (R_B^E)^{-1} R_S^E R_C^S \quad (10)$$

A planar constraint on a set of points $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$, and with normal vector $\mathbf{n}$, is written as a set of $m-1$ constraint equations, each having the form:

$$\mathbf{n}^T R_B^E (\mathbf{p}_j - \mathbf{p}_{j+1}) = 0, \quad 1 \le j < m \quad (11)$$

2.5. GPS as Positional Constraints

PALM exploits the linear formulation in using the GPS constraints. From (1) and (2), it is clear that camera translation is coupled with shape. If knowledge of the camera translations is available through GPS measurements, the overall shape can be constrained accordingly, using (6). The advantage of using GPS is that the errors do not propagate from point to point.

2.6. Avoiding trivial solutions

In solving for the complete shape vector $\mathbf{x}$, two trivial solutions exist. The first solution is to set $\mathbf{x}$ to be the zero vector, which obviously satisfies (7) and (11). The second is to set all points $\mathbf{p}_p$ and camera translations $\mathbf{t}_f$ to be identical and equal to an arbitrary 3-vector. In this case, (11) is clearly satisfied. Since (7) is derived from (3) and (4), it is satisfied as well.

To prevent these trivial solutions, one of two approaches can be used. If GPS readings are available, (6) can be used to constrain the camera locations. If GPS readings are not available, two points from the complete large structure are picked and their distance set to a non-zero value.

2.7. The Constraint-based Solver

The PALM solution method consists of two solvers: linear and non-linear. The linear solver provides initial solution estimates which serve as input to the non-linear solver.

2.7.1 The Linear Solver

The linear system of equations is formed by combining (7) for all frames f, (11) for all planes, and (6) if GPS readings are available. This linear system is used to solve for the complete shape vector $\mathbf{x}$.

Two solution methods were tried: hard/soft constraints model [5]; and LU decomposition. It was found that both gave good estimates as initial solutions to the non-linear optimizer.

2.7.2 The Non-Linear Solver

The non-linear solver is implemented using the Levenberg Marquardt technique. This optimization refines all the estimates, including all the shape points $\mathbf{p}_p$, all camera translations $\mathbf{t}_f$, all camera orientation matrices $[\mathbf{i}_f \ \mathbf{j}_f \ \mathbf{k}_f]^T$, and all building orientations with respect to earth frame R_B^E. Quaternions are used to represent all rotations.

The energy function to be minimized is

$$E = E_{point} + \alpha E_{planar} + \beta E_{gps} \qquad (12)$$

In our experiments, we set α to be 1 and β to be a low value (0.0001). The GPS energy term is not emphasized in this non-linear refinement stage because the main use of camera positional constraints is to ensure a good overall shape in the linear solver stage. GPS measurements contain errors that may distort the non-linear refinement process.

E_{point} is the total projection error for all feature points in all frames and it is given by

$$E_{point} = \sum_f \sum_p [(u_{fp} - \frac{l(\mathbf{p}_p - \mathbf{t}_f) \cdot \mathbf{i}_f}{(\mathbf{p}_p - \mathbf{t}_f) \cdot \mathbf{k}_f})^2 + (v_{fp} - \frac{l(\mathbf{p}_p - \mathbf{t}_f) \cdot \mathbf{j}_f}{(\mathbf{p}_p - \mathbf{t}_f) \cdot \mathbf{k}_f})^2] \qquad (13)$$

E_{planar} is the sum of errors caused by deviation of points from their assigned planes. For each constraint plane,

$$E_{one\_plane} = \sum_{j=1}^{m-1} [\mathbf{n}^T R_B^E (\mathbf{p}_j - \mathbf{p}_{j+1})] \qquad (14)$$

where $\mathbf{n}$ is the plane normal and m is the number of points on the plane.

Note that $\mathbf{n}$ is defined local to the object frame. For scenes with multiple objects, $\mathbf{n}$ for each object need to be transformed to the global frame through the matrix R_B^E. R_B^E can be estimated using a view of the building that contains pairs of horizontal and vertical lines.

If GPS readings are available, they can be used to constrain the camera translations using

$$E_{gps} = \sum_{f \in \Omega} (\mathbf{t}_f - \mathbf{g}_f)^T (\mathbf{t}_f - \mathbf{g}_f) \qquad (15)$$

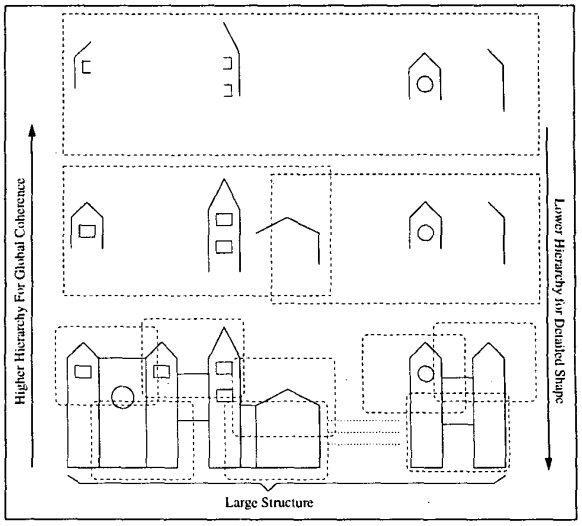


Figure 3. Shape Hierarchy: dotted boxes represent independent views in each of the hierarchies

Ω is the set of all frames where GPS readings are available, and $\mathbf{g}_f$ is the GPS reading at frame f.

2.8. Reduction of Merging Errors

PALM uses a technique named Landmarking to alleviate the merging error problem in the reconstruction of a large scene.

The idea of landmarking is to seek out a few camera views, each of which sees more than one of the smaller shape segments. In each of these landmark views, several points are selected. These landmark points are matched with the corresponding points in the relevant shape segments. Conceptually, landmark points project to rays in the 3D space to constrain the relative positioning of the shape segments. These ray constraints are written using (7).

Landmarking for arbitrary scenes (i.e., structured, unstructured, or a combination of both) is feasible in PALM because of the use of the heading/tilt sensor. The key idea here is that the camera orientation readings make it possible to enforce the landmark constraints from as little as one landmark image. While structure-from-motion requires multiple images taken with large camera translations, this requirement is not necessary for landmarking. This is an important property because landmarked areas are typically bigger and possibly have a depth larger than the object-to-camera distance, and so conventional structure-from-motion techniques will likely give poor accuracies.

Landmarking has two other properties that are also of practical importance. As little as one landmark point on

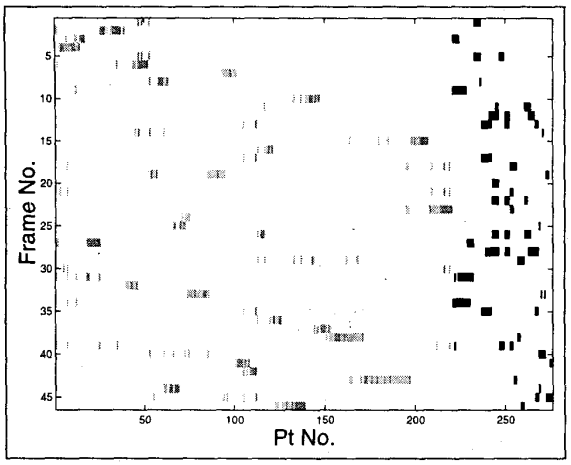


Figure 4. Observation map of feature points for the stadium model. Gray pixels represent observed points belonging to planes. Dark pixels represent observed points that do not belong to planes or any other geometrical objects. Empty spaces represent occlusion.

each shape segment is useful and not all shape segments need to contain landmark points. These properties help in the overall shape reconstruction because large structures usually consist of parts that occlude each other, so views that contain big portions of the structure are likely to see only partial views of the shape segments. Fig. 3 shows the decomposition of a large structure into shape hierarchies. Each dotted box represents an independent view. At low levels in the hierarchy, local but detailed views are captured; at high levels, information on the overall shape is available from the views. It should be noted that landmarking deals with images at high levels in the shape hierarchy.

2.9. Small Number of Images and Features

Structure-from-motion techniques require point correspondences across image frames. Most structure-from-motion methods require the feature points to be observed in many frames. This requirement is not true in PALM. PALM requires only a small number of images and feature points. Because of the use of the camera heading/tilt sensor, the complete scene reconstruction can be solved as a single linear system even if the observation map is sparse. Fig. 4 is an example of the sparse observation map for the stadium model (Section 3.3). Each point in the map is observed in a relatively small number of frames.

If a scene contains special features like points on known planar orientation, the number of images required can be further reduced because the 3D coordinates of these planar points can be recovered from just one image.

3. Experiments

We used PALM to reconstruct two large buildings and a football stadium in a campus environment. The plan view dimensions of these large structures are 425x164 ft, 434x351 ft, and 716x486 ft, respectively.

The effectiveness of landmarking is illustrated in all three examples. The third example – the football stadium, illustrates the flexibility of PALM in dealing with scenes that comprise both structured and unstructured shapes.

The images used are partial views of the scene and were taken at ground level; no area views are used.

The plan views of these stadium/buildings are digitized from the architectural blueprints for comparison (Fig. 5,11,14). Circular marks in these figures indicate the ground truth points that are used to evaluate our results.

3.1. Building 1: 425x164 ft

Fig. 5 shows the ground truth plan view of Building 1. The dotted lines correspond to a portion of the structure that is not modeled in our experiments. Therefore, no ground truth points are chosen in that region.

Images of the building were captured and the entire building was broken up into 14 shape segments.

Without the use of landmark and GPS constraints, the complete shape reconstructed is as shown in Fig. 6(a). The shape segments are totally out of scale for the left (enclosed by a circle) and the right sections of the building. The huge scaling error is due to the fact that the merging was forced to take place at a narrow region because of occlusion by the part of the structure that was not being modeled (Figs. 6(b),6(c)). The shape segments were forced to be merged by points at close distance. As a result, the relative scaling calculation became unstable. Shape reconstruction errors within a shape segment were multiplied with the feature location errors at the merging. This explains the large scaling error as shown in Fig. 6(a).

Our landmarking technique resolves this problem. Three landmark images were taken for this building. One of the landmark images is as shown in Fig. 7, with twelve landmark points.

The solution using landmark constraints (without GPS) is shown in Fig. 8(a). The peak error in this case is 15 ft (the dotted curve in Fig. 10).

Next, we experimented with the GPS constraints. GPS readings were taken at eleven of the fourteen camera locations. When these GPS constraints were incorporated, the scaling problem shown in Fig. 6 was resolved, even though

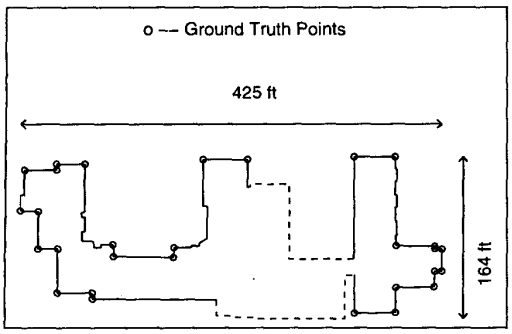


Figure 5. Plan View of Building 1

no landmark constraints were used. The shape reconstruction results using GPS and without landmark constraints is shown in Fig. 8(b). Notice that the large scaling error has disappeared. This is not surprising because, as explained in Section 2.5, camera translations are sources of shape information in the perspective projection model, and so by enforcing the correct values for camera translations, the reconstructed shape will be close to the correct shape. The peak error in this case was 23 ft (Fig. 10).

Fig. 8(c) shows the solution using both landmark and GPS constraints, and Fig. 9 shows the recovered camera locations. The peak error in this case was 7 ft (Fig. 10). The GPS readings effectively reduced the peak error from 15 ft (when only landmark constraints were used) to 7 ft (when both landmark and GPS constraints were used).

3.2 Building 2: 434x351 ft

The ground truth points for Building 2 are shown in Fig. 11. The building is broken up into 16 shape segments. Landmark constraints were used, but GPS readings were not used for this experiment.

The complete shape reconstructed without using landmark constraints is shown in Fig. 12(a). One can notice that the reconstructed building has its two protruding portions misaligned, as indicated by the arrows.

This misalignment error was due to the fact that the plane (indicated by the arrow in Fig. 12(b)) was viewed from a direction such that its normal vector was almost perpendicular to the camera optical axis. A small error in feature location induced huge errors in the reconstruction.

Again, the landmark technique effectively fixed this problem. The landmark points are as shown in Fig. 12(c). Fig. 13(a) shows the solution after the non-linear optimization stage , and Fig. 13(b) shows the recovered camera locations. The peak error was 17 ft (Fig. 13(c)).

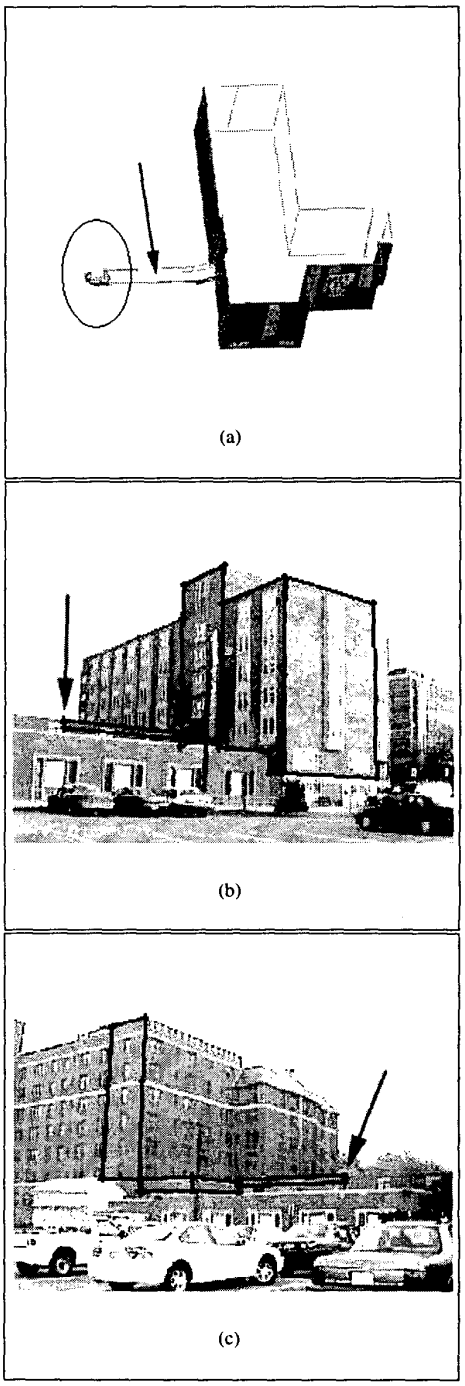


Figure 6. Large scaling error that occurs when merging takes place at a narrow region (arrows point to location of merge). Top: Reconstructed model, left and right portion out of scale; Middle and bottom: Images used for merging.

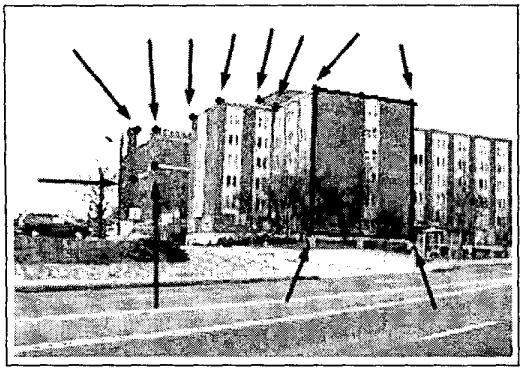


Figure 7. Landmark points (Building 1)

3.3. Stadium: 716x486 ft

This example is difficult compared with the previous two because:

1. The scene consists of 3 unrelated and disjointed buildings. These buildings do not share any features or objects that constrain relative locations and orientation.
2. The images were taken from the stadium field, thus they are "looking out" at the scene being reconstructed. This means relatively shorter baselines compared with those of the first two experiments in which the paths traced by the camera were longer than the perimenter of the buildings.

To relate each of the building orientations with respect to the earth frame, one view from each building that contains pairs of horizontal and vertical lines is selected. These horizontal and vertical lines are used to estimate the camera orientation (R_C^B) with respect to each of the building frames. Since camera orientation ($R_C^E = R_S^E R_C^S$) with respect to earth is given by the orientation sensor, the building orientation (R_B^E) with respect to earth frame can be estimated using (9).

For the unstructured space in between buildings, end points on the lamp posts are chosen to be recovered.

Without landmarking and GPS constraints, the reconstructed shape would have been as shown in Fig. 15. Notice that the reconstructed shape as well as the recovered camera positions have a huge error. When landmarking and GPS were used, the reconstructed shape and camera pose were much better as shown in Fig. 17(a). Figs. 17(b),17(c) show two views of the reconstructed model.

4. Conclusions

We proposed PALM that addresses the merging error problem in large scene reconstruction. PALM is

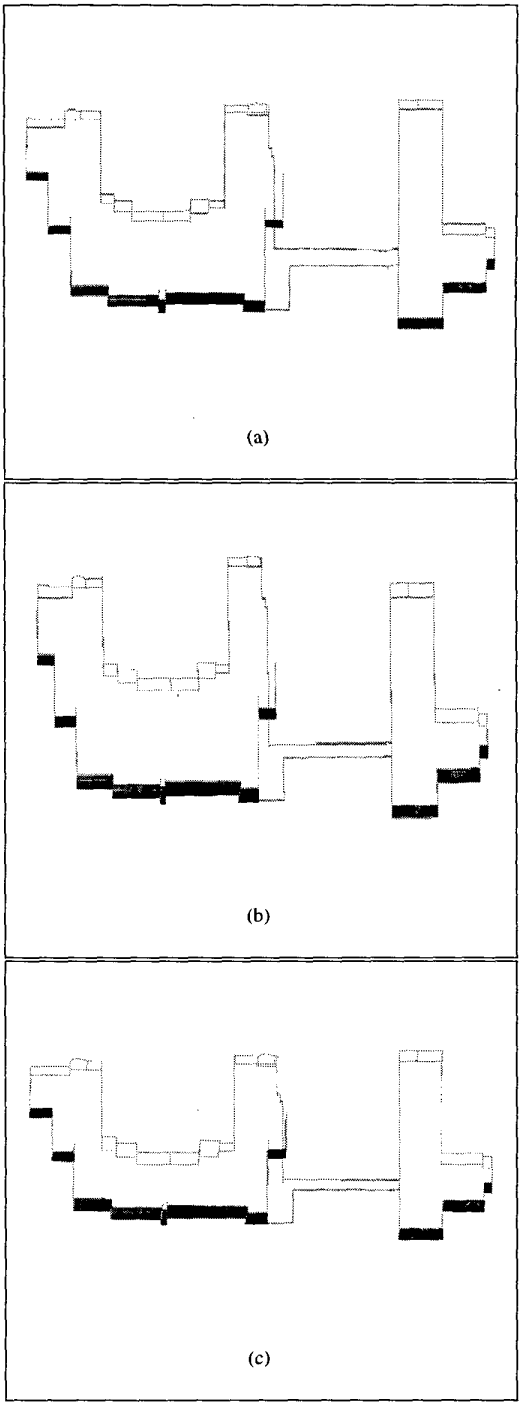


Figure 8. (a) Recovered shape using landmark constraints (without GPS). (b) Recovered shape using GPS constraints (without landmarking). (c) Final recovered shape using landmark and GPS constraints

Figure 9. Recovered Build. 1 and camera loc.

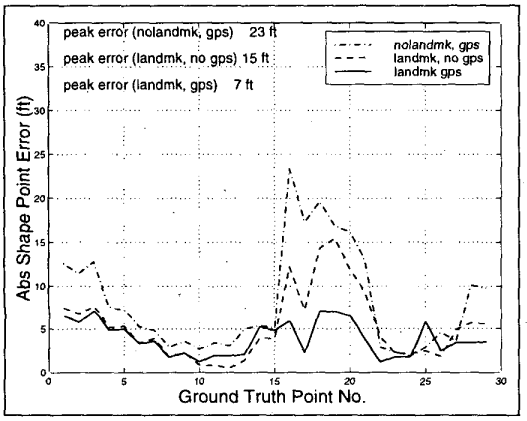


Figure 10. Comparison of Shape Error

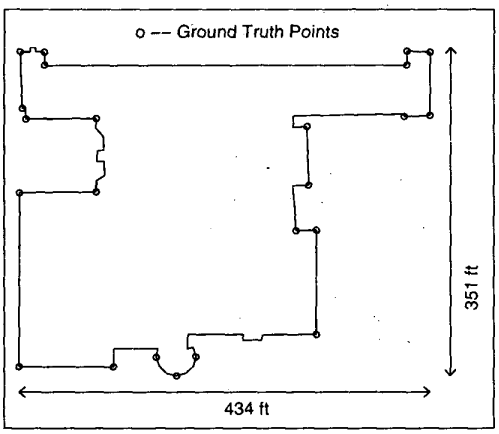


Figure 11. Plan View of Building 2

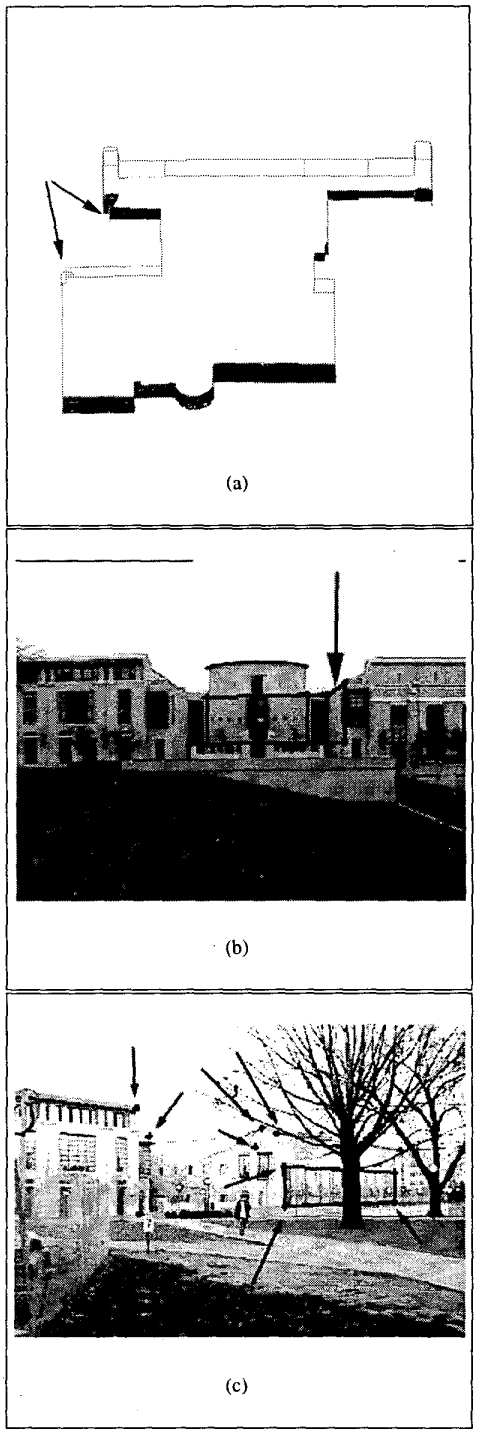


Figure 12. (a) Two portions misaligned in the reconstructed shape. (b) Cause of the misalignment: plane normal almost perp. to optical axis. (c) Landmark points used to fix the misalignment problem

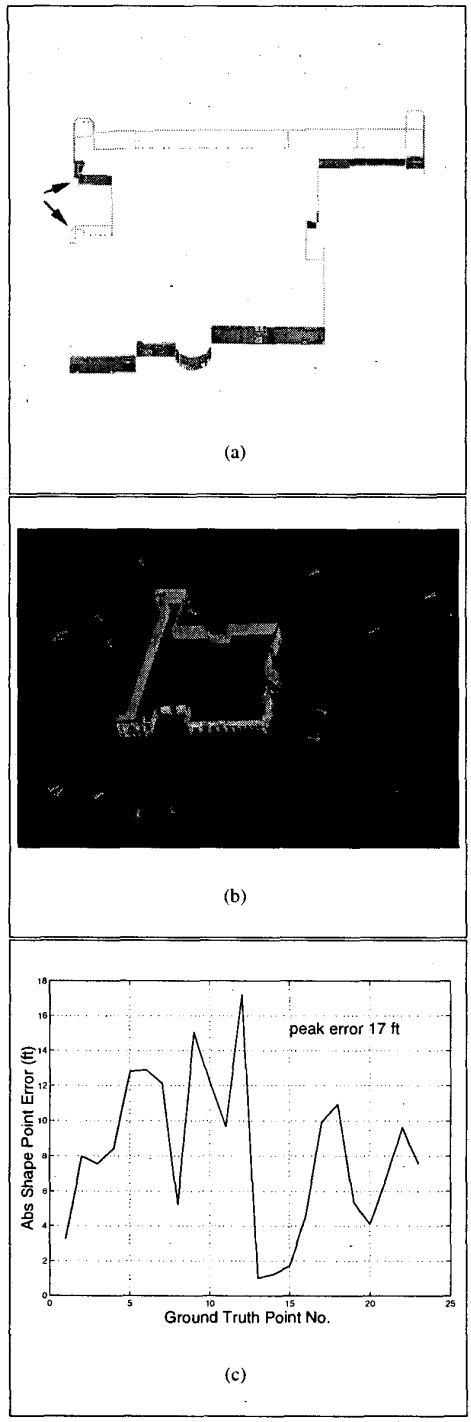


Figure 13. (a) Final reconstructed shape using landmark constraints: misalignment reduced (b) Reconstructed Building 2 and Camera Locations (c) Shape error

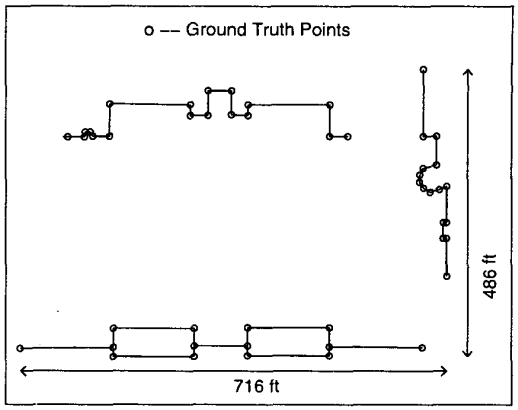


Figure 14. Plan View of Stadium

Figure 15. Huge errors occur in the recovered shape and camera positions if landmarking and GPS are not used.

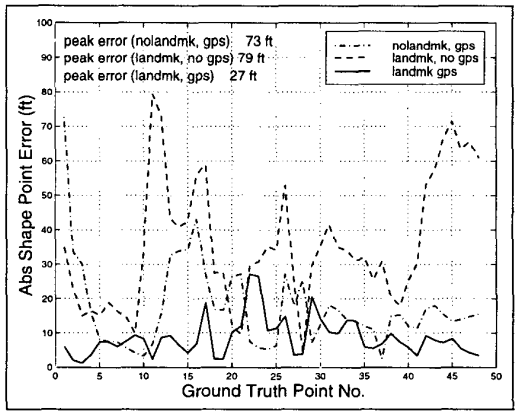
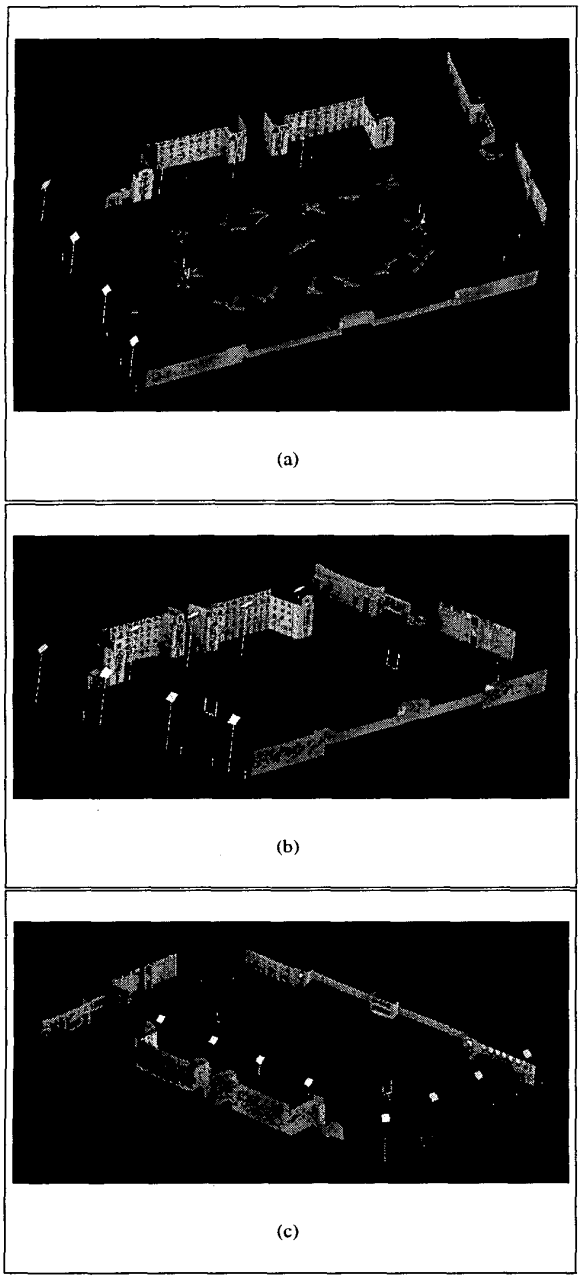


Figure 16. Comparison of Shape Error



(a)

(b)

(c)

Figure 17. Reconstructed stadium using landmark and GPS constraints. (a) Reconstructed stadium and camera pose (b) A view of the reconstructed stadium (c) Another view of the reconstructed stadium

equipped with a camera heading/tilt sensor that helps to constrain the overall shape, regardless of whether the scene is structured or unstructured. The merging error problem was solved by the landmarking technique. Comparing the results with the ground truth, we conclude that landmarking and GPS improve the accuracy. The results for each of the reconstructions of the stadium/buildings are summarized in the following table:

	length	width	max error (ft)
Build1	425	164	15
Build1(gps)	425	164	7
Build2	434	351	17
Stadium	716	486	27

Table 1: Maximum Absolute Shape Error

Acknowledgements

We thank Toshihiko Suzuki for contributing the circuit design of the hardware encoder; David LaRose for calibrating our camera using his very efficient camera calibration method; Mei Han for bringing to our attention the concept of hard and soft constraints, and the tremendous help all three have rendered. We also thank Dave Duggins and Bob Collins for helping out with the GPS measurements, and Marie Elm for her editorial help. The support of DTG, Singapore, is also greatly appreciated.

References

[1] A. Azarbayejani and A. Pentland. Recursive estimation of motion, structure and focal length. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(6):562–575, June 1995.

[2] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. *Technical Report UCB//CSD-96-893, University of California at Berkeley*, January 1996.

[3] D. LaRose. A fast, affordable system for augmented reality. *Technical Report, Carnegie Mellon University, CMU-RI-TR-98-21*, April 1998.

[4] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3:209–236, 1989.

[5] H. Shum, M. Han, and R. Szeliski. Interactive construction of 3d models from panoramic mosaics. *In Proceedings of the Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, USA.*, pages 427–433, 1998.

[6] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, November 1992.