Automatic 3D Modelling of Architecture

Anthony Dick[1] Phil Torr[2] Roberto Cipolla[1]

[1] Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK

[2] Microsoft Research, 1 Guildhall St, Cambridge CB2 3NH, UK

ard28@eng.cam.ac.uk philtorr@microsoft.com cipolla@eng.cam.ac.uk

Abstract

This paper describes a system which automatically derives 3D models of architectural scenes from multiple images. This system differs from previous structure from motion algorithms in that it explicitly makes use of strong geometric constraints such as perpendicularity and verticality which are likely to be found in architecture. Structure is compactly represented as a piecewise planar model which is initialised automatically by segmenting a feature-based reconstruction. An efficient technique for evaluation of model likelihood is also presented, which allows a rapid search through a large number of 3D models.

1 Introduction

As 3D graphics become an everyday feature of desktop PCs and web browsers, there is an increasing demand for the generation of realistic 3D models without the use of expensive and technical CAD packages. The automatic acquisition of 3D models from image sequences is one approach to the problem which has been actively pursued by computer vision researchers in recent years.

Fully automatic model recovery from images [3, 9] is usually based on low level image features such as points and lines, which can be automatically detected fairly reliably. However reconstructing each point or line individually involves estimating a large number of parameters ($3n$ parameters for n points, for example), which makes this approach ill-conditioned and reliant on many accurately detected and matched point and line features.

By relaxing the requirement for complete automation, interactive systems [4, 10, 11] can make use of higher level primitives such as planes and blocks, which are manually extracted from an image sequence. This greatly reduces the dimensionality of the problem and leads to more robust solutions and convincing models. However these systems can only model scenes in which the primitives they use are applicable, and are no longer completely automatic.

This paper describes a system for automatic 3D model acquisition which makes use of high level primitives (planes). The system is demonstrated on architectural scenes, although it could be applied to any class of structure for which some prior knowledge is available. From an initial reconstruction obtained from corner features, a model based on planes, similar to the layered models used in [2, 13], is initialised and then optimised. This process explicitly makes use of strong prior models and geometric constraints which are applicable to architecture. This has the same aim as that of Baillard et. al.[1]; however their work follows a different approach based on line matching to segment planes from aerial images of urban scenes.

Section 2 of the paper introduces the piecewise planar model used to represent architectural scenes, and defines a posterior probability measure for the model. Section 3 addresses the problem of obtaining an initial estimate of the model automatically from an image sequence, while in Section 4 a method for rapidly optimising this estimate is presented. In Section 5 the model is further refined by the addition of offset layers representing features such as doors and windows to each plane.

2 The planar model

Each plane p in a planar model is specified by spatial parameters $\boldsymbol{\theta}_S^p$, boundary parameters $\boldsymbol{\theta}_B^p$, texture parameters $\boldsymbol{\theta}_T^p$ and parallax parameters $\boldsymbol{\theta}_Z^p$. The spatial parameters $\boldsymbol{\theta}_S^p$ determine the position and orientation of plane p in the $[XYZ]$ world coordinate system and are defined such that $\boldsymbol{\theta}_S^p \cdot [X\ Y\ Z] = 1$. The boundary parameters $\boldsymbol{\theta}_B^p$ are a list of points which defines a clockwise walk around the boundary of the plane. The last point in $\boldsymbol{\theta}_B^p$ is joined to the first to form a closed curve, which is assumed not to be self-intersecting. The texture parameters $\boldsymbol{\theta}_T^p$ consist of an unknown brightness parameter $i(\mathbf{X})$ (between 0 and 255) defined at each point $\mathbf{X}$ on a regular 2D grid within the plane boundary. Also defined at each point $\mathbf{X}$ is a parallax parameter $z(\mathbf{X})$, which models the offset of point $\mathbf{X}$ perpendicular to the plane. The set of offset parameters forms the parallax parameter vector $\boldsymbol{\theta}_Z^p$. The planar model also contains a background plane, p_∞, with infinite extent.

2.1 Estimating a planar model

Given data $\mathbf{D}$ in the form of an image sequence, the problem is to find $\boldsymbol{\theta} = \bigcup_p \boldsymbol{\theta}_S^p \boldsymbol{\theta}_B^p \boldsymbol{\theta}_T^p \boldsymbol{\theta}_Z^p$ such that the posterior probability

$$p(\boldsymbol{\theta}|\mathbf{DI}) \sim p(\mathbf{D}|\boldsymbol{\theta}\mathbf{I})p(\boldsymbol{\theta}|\mathbf{I}) \tag{1}$$

is maximised, where $\mathbf{I}$ represents prior information. This is a product of model likelihood and prior, which are defined in the following sections.

2.2 Evaluation of the model likelihood

It is assumed that a texture parameter at point $\mathbf{X}$ on the model surface is observed with noise $i(\mathbf{X}) + \epsilon$, where ϵ has a Gaussian distribution mean zero and standard deviation σ_ϵ. The parameters $i(\mathbf{X})$ and $z(\mathbf{X})$ can then be found such that they minimise the sum of squares $\sum_{j=1}^{j=m} \left(i(\mathbf{x}^j) - i(\mathbf{X})\right)^2$ where $i(\mathbf{x}^j)$ is the intensity at $\mathbf{x}^j$, and $\mathbf{x}^j$ is the projection of $\mathbf{X}$ with offset $z(\mathbf{X})$ into the jth image. The likelihood for given values of $i(\mathbf{X})$ and $z(\mathbf{X})$ is

$$p(\mathbf{D}|i(\mathbf{X})z(\mathbf{X})) = \left(\sqrt{2\pi}\sigma_\epsilon\right)^{-j} \exp -\frac{1}{2}\sum_j \left(\frac{i(\mathbf{x}^j) - i(\mathbf{X})}{\sigma_\epsilon}\right)^2 \tag{2}$$

Using Equation (2) under the assumption that the errors ϵ in all the points are independent, the likelihood over all points can be written:

$$p(\mathbf{D}|\boldsymbol{\theta}\mathbf{I}) = \left(\sqrt{2\pi}\sigma_\epsilon\right)^{-ij} \exp -\frac{1}{2}\sum_i \sum_j \left(\frac{i(\mathbf{x}_i^j) - i(\mathbf{X}_i)}{\sigma_\epsilon}\right)^2 \tag{3}$$

where $\mathbf{x}_i^j$ is the projection of the ith scene point $\mathbf{X}_i$ into the jth image.

2.3 Evaluation of the prior

Assuming the prior distributions of $\boldsymbol{\theta}_S$, $\boldsymbol{\theta}_B$, $\boldsymbol{\theta}_T$, and $\boldsymbol{\theta}_Z$ are independent, the prior $p(\boldsymbol{\theta}|\mathbf{I})$ is factorised as $p(\boldsymbol{\theta}_S|\mathbf{I})p(\boldsymbol{\theta}_B|\mathbf{I})p(\boldsymbol{\theta}_T|\mathbf{I})p(\boldsymbol{\theta}_Z|\mathbf{I})$. Given no prior knowledge of the appearance of a model or the parallax from its planes, both $p(\boldsymbol{\theta}_T|\mathbf{I})$ and $p(\boldsymbol{\theta}_Z|\mathbf{I})$ are assigned uniform distributions. Assigning a distribution to plane shape, i.e. $p(\boldsymbol{\theta}_B|\mathbf{I})$, is an interesting problem, but is not tackled in this paper. For now only a simple prior is applied to the spatial parameters $\boldsymbol{\theta}_S$: it is expected that adjacent walls are likely to intersect at about 90 degrees. A simple way of encoding this is to define a Gaussian prior, centred at $\pi/2$ and with standard deviation σ_ϕ, on the angle between adjacent walls:

$$p(\boldsymbol{\theta}|\mathbf{I}) = \left(\sqrt{2\pi}\sigma_\phi\right)^{-(n-1)} \exp -\frac{1}{2}\sum_{i=1}^{n-1} \left(\frac{\phi_i - \pi/2}{\sigma_\phi}\right)^2 \tag{4}$$

where n is the number of planes in the model and ϕ_i is the interior angle between planes i and $i+1$ when projected onto a ground plane (see Section 3).

Having defined the components of the posterior probability measure, the maximum a posteriori (MAP) parameter vector $\boldsymbol{\theta}_{\text{MAP}}$ is now sought. There are two stages to the search: first a good initial estimate is obtained, and then this estimate is refined using a search algorithm.

3 Initialising the planar model

The initial estimate of the planar model is obtained from a feature-based 3D reconstruction. This section briefly describes the acquisition of the feature-based 3D model, which is based largely on previous work [3, 9], and how it is segmented to derive an initial estimate of the planar model.

3.1 Point matching in architectural scenes

The feature-based model is based on image corners, which are detected using a Harris corner detector [6] with subpixel interpolation. For each corner at position (x, y) in the first image, candidate matches from a neighbourhood of (x, y) in the second image are ranked by cross-correlation. To reduce the number of mismatches, the epipolar geometry (i.e. the fundamental matrix $\mathbf{F}$) between the image pair is often estimated using a technique which is tolerant to outliers, such as RANSAC[5]. However this was found to be error-prone for architectural scenes such as that in Figure 1, which contains significant amounts of repeated structure, and planar surfaces. A better strategy in this case is to robustly estimate a planar homography $\mathbf{H}$ from the initial correspondences. Although not all correct matches are consistent with $\mathbf{H}$, this resolves much of the ambiguity associated with planar repeated structure, as shown in Figure 1. Correct matches which are not consistent with $\mathbf{H}$ are displaced by a residual parallax vector. These parallax vectors intersect at the epipole $\mathbf{e}_2$ in the second image; hence the epipole is obtained by a robust intersection of parallax vectors. The epipolar geometry between the first two images is now fully determined, and the associated fundamental matrix is given by

$$\mathbf{F} = [\mathbf{e}_2]_\times \mathbf{H} \tag{5}$$

where $[\mathbf{e}_2]_\times$ is the skew-symmetric matrix formed from $\mathbf{e}_2$. More reliable matching is now carried out by searching for matches along corresponding epipolar lines.

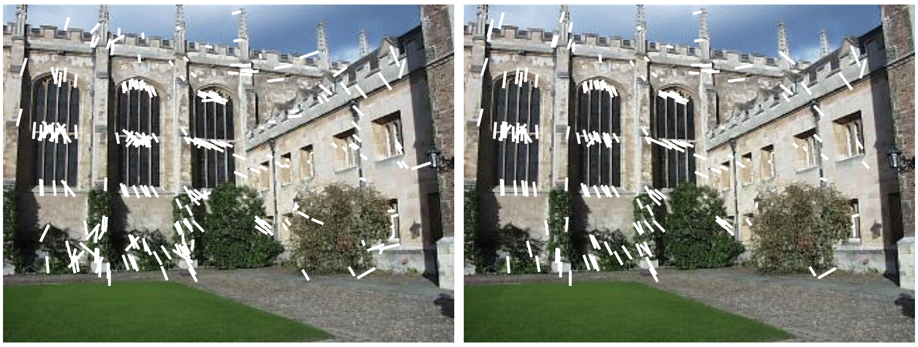


Figure 1: *Correspondences obtained from* $\mathbf{F}$ *estimated directly from the data (left) and* $\mathbf{F}$ *estimated from a dominant plane plus parallax (right). The correspondences obtained via plane plus parallax are more reliable, particularly where there is repeated structure (e.g. in the chapel windows).*

3.2 Camera Calibration

Assuming that the camera has an aspect ratio near 1, and a principal point near the centre of the image, only the focal length is required to calibrate the cameras sufficiently to produce an accurate reconstruction. The essential matrix between the first two cameras, $\mathbf{E} = \mathbf{K}_1^T \mathbf{F} \mathbf{K}_2$, where $\mathbf{K}_1$ and $\mathbf{K}_2$ contain the intrinsic parameters of cameras 1 and 2 respectively, has two equal eigenvalues σ_1 and σ_2. In a simplified version of [8], the penalty function

$$\mathcal{C} = \frac{|\sigma_1 - \sigma_2|}{\sigma_2} \tag{6}$$

is minimised by direct search, varying the focal length of each camera in turn. Having obtained correspondences and calibration, a cloud of 3D points is triangulated, and a bundle adjustment routine is used to minimise the reprojection error.

3.3 Incorporating matches from other images

Other images are now sequentially incorporated into the reconstruction. Each image is matched with its predecessor as described in Section 3.1, which provides both 2D-2D correspondence and 2D-3D correspondence, as points in the previous image already have associated structure. A projection matrix is then estimated linearly from the 2D-3D correspondences, and new structure is triangulated from 2D-2D matches without previous structure. After each image is incorporated, the reconstruction and projection matrices are again optimised using bundle adjustment. A model obtained using this method is shown in Figure 2.

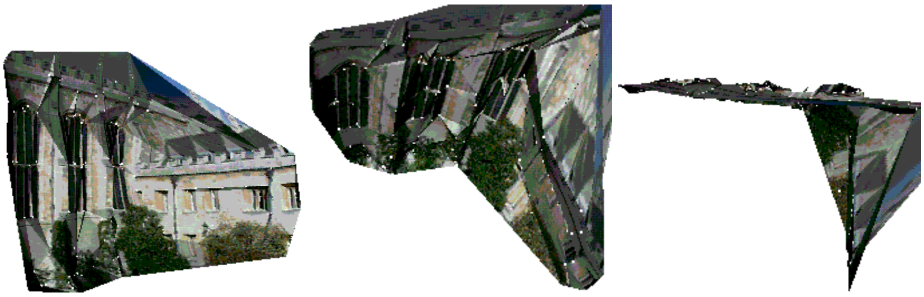


Figure 2: *Reconstruction obtained from triangulated points only. The points are joined by triangular texture mapped faces, obtained using Delaunay triangulation. This reconstruction is quite noisy.*

3.4 Extracting planes from the reconstruction

Planes are extracted from the 3D reconstruction by recursively applying a RANSAC plane estimation algorithm. At each iteration, planes are hypothesised from a randomly selected sample of 3 points, and the plane which is best supported by those points which have not been assigned to previous planes is selected.

The main problem with this approach is that it includes no concept of the spatial extent of each plane, and thus tends to extract planes with irregular and mutually overlapping boundaries. To impose spatial constraints it is assumed that all planes are perpendicular to a common ground plane—an assumption which is justified for architectural scenes in which dominant planes commonly correspond to walls of buildings. Provided that at least two planes have been extracted from the point model, a least squares estimate of the ground plane normal can be obtained from the other plane normals. An overhead view of the reconstruction can now be simulated by projecting all points orthogonally onto the ground plane, as shown in Figure 3. This view is useful as many strong constraints applicable to architecture can be applied to it—in fact it is to this view that the perpendicularity prior (Equation (4)) is applied. Also shown in Figure 3 are two lines which have been fitted to the inlier clusters obtained from the recursive RANSAC estimation. These lines are intersected, which provides an initial guess of the spatial extent (parallel to the ground plane) of each plane. The upper and lower bounds of each plane are also initialised by the ”highest” and ”lowest” points in each cluster, where height is measured perpendicular to the ground plane.

Thus a simple initial planar model is obtained in which all planes have rectangular shape and are perpendicular to a ground plane. This model is next used to seed a search for an improved estimate.

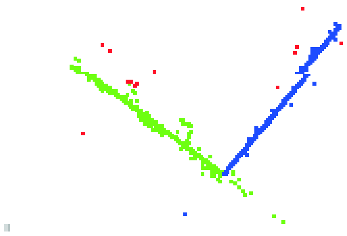
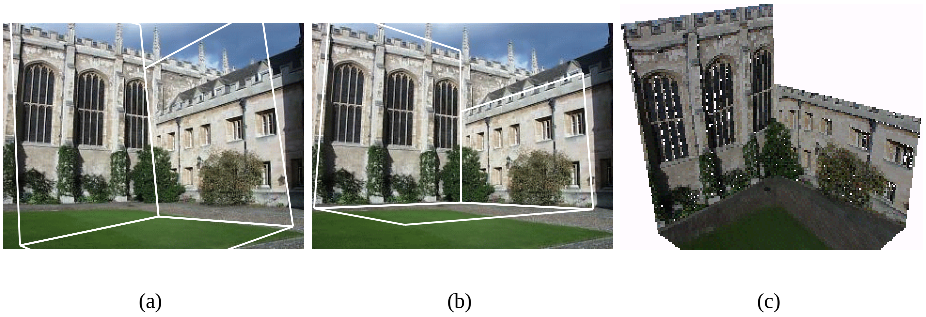


Figure 3: *Overhead view of the reconstruction. Green (lighter grey) points belong to the left plane; blue (darker) points belong to the right. Outliers are shown in red. Fitted lines are superimposed on the view.*

4 Optimising the planar model

Having initialised a 3D model consisting of a set of rectangles, and defined a probability measure for the model, a search for the MAP model parameters is now carried out. There are three main sources of error in the initialisation process described in Section 3.4: the least squares estimate of the ground plane normal, the segmentation of the planes by intersecting their projections on the ground plane, and the approximation of each plane boundary by a rectangle. Hence the model is optimised by gradient descent search on (a) the components of the ground plane normal, (b) the endpoints of the projection of each plane onto the ground plane and (c) the boundary of each plane.

Projections of the planar model are shown before and after searching in Figure 4, along with a texture mapped version of the resultant 3D model. The search typically takes a total of order 50 iterations to converge, which takes about 20 seconds on a 500MHz Pentium 3. The speed of the algorithm is mainly due to the technique now presented, which makes evaluation of the model likelihood extremely fast.



(a) (b) (c)

Figure 4: *Projection of the planar model (a) before and (b) after searching. The negative log likelihood of model (b) is* 1.82×10^5*, compared to* 1.93×10^5 *for the unoptimised model (a). The texture mapped planes are shown in (c).*

4.1 Efficient computation of the model likelihood

Computing the exponent of the likelihood equation (Equation (3)) involves integrating the sum-squared intensity error (Equation (2)) over a grid of points on the surface of each plane. As in [7], this integral over a region $\mathcal{R} \subset \mathbf{R}^2$ is rewritten as an integral around the region boundary $\partial\mathcal{R}$ using Green's Theorem:

$$\int_{\mathcal{R}} f(x,y)dxdy = \int_{\partial\mathcal{R}} \mathbf{n} \cdot \vec{A} ds \tag{7}$$

where $f(x,y)$ is a real-valued function of $\mathbf{R}^2$, $\mathbf{n}$ is the unit normal to the boundary and $\vec{\nabla} \cdot \vec{A} = f$. For evaluation of the likelihood of plane p, $f(x,y)$ is $p(\mathbf{D}|i(\mathbf{X})z(\mathbf{X}))$, and $\partial\mathcal{R}$ is given by the boundary parameters $\boldsymbol{\theta}_B^p$. The vector field $\vec{A}(x,y)$ is given by

$$\vec{A}(x,y) = \frac{1}{2} \begin{bmatrix} \int_0^x f(\tilde{x},y)d\tilde{x} \\ \int_0^y f(x,\tilde{y})d\tilde{y} \end{bmatrix} \tag{8}$$

Hence the likelihood of plane p in a region $\mathcal{R}$ is found by computing $\mathbf{n} \cdot \vec{A}$ for each pixel lying on a discretised version of $\partial\mathcal{R}$. At each pixel $\mathbf{u}$ on the boundary, this is approximated by

$$\det \left(\begin{bmatrix} (\mathbf{v} - \mathbf{u})^T \\ \left[\vec{A}(\mathbf{u}) + \vec{A}(\mathbf{v})\right]/2 \end{bmatrix} \right) \tag{9}$$

where $\mathbf{v}$ is the successor pixel to $\mathbf{u}$ on $\partial\mathcal{R}$, in an anti-clockwise direction. This is equivalent to computing $\mathbf{n} \cdot \vec{A}$ for each line segment joining adjacent pixels, with a weighting factor proportional to the length of this line segment.

To compute the likelihood of the entire model, Equation (9) is evaluated around the boundary of each plane, including the background plane. It is assumed that all planes except the background are simply connected. Clearly the background plane is not simply connected—each of the other planes lies inside its boundary. Hence the likelihood of the background plane must be calculated around the boundary of each region and subtracted from the total model likelihood (or equivalently, the background likelihood must be integrated clockwise around each region boundary). This is shown graphically in Figure 5 (a). By precomputing the vector field $\vec{A}$ for each surface, calculation of the likelihood only involves determining the boundary pixels for each region and summing these precomputed values over the region boundaries, which is very fast. If the results of the integration are cached for each line segment belonging to each boundary, recalculation of the likelihood for a slightly altered model is even faster, requiring only the recalculation of the affected boundary segments, as shown in Figure 5(b).

5 Model refinement using planar parallax

The planar model shown in Figure 4 is less noisy than the point-based model given in Figure 2, but it is missing a lot of detail. Because each wall is modelled as a single plane, details such as doors, windows and pillars are excluded from the model. Recall that the likelihood of each point $\mathbf{X}$ on the model surface is maximised over a small range of offset values $z(\mathbf{X})$. In theory, a region growing algorithm could be used to search for

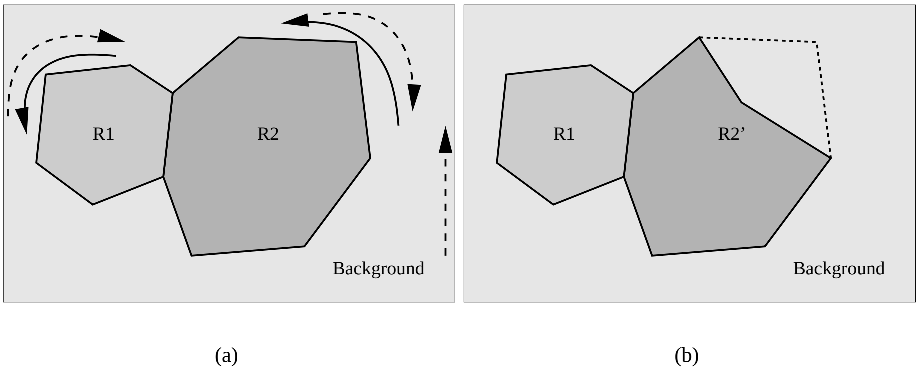


Figure 5: *(a) Likelihood evaluation using Green's Theorem. Rather than computing the likelihood (Equation (3)) in regions R1, R2 and the background, a related vector field is integrated anticlockwise around the boundary of each region (solid arrows). The vector field associated with the background is also integrated clockwise around the boundaries of regions interior to it (dashed arrows). (b) Rapid evaluation of likelihood for varying shapes. Having computed the likelihood of region R2, the likelihood of region R2$'$ is computed by subtracting the contributions from the invalid boundary segments (shown as dashed lines) and adding those of the new boundary segments.*

large spatial regions within a plane whose likelihood is maximised at similar offset values. However there is generally insufficient detail in images of multiple walls to reliably recover the position and number of offset layers this way. In future means of merging distant images of multiple planes and close-up images of single planes will be investigated; however at present the approximate position and size of offset layers is manually initialised. Once these layers have been initialised, their shape is automatically selected from a predefined family of parameterised shape models as described in [12]. This family includes shapes which correspond to commonly occurring architectural features, such as a rectangle, an arch and a pillar. For instance the arch is represented by 8 parameters: position on plane (x, y), a width and height a and b, an offset from the plane d, orientation ω, bevel r (the slope of the edges) and arch height c. A Bayesian model selection technique is used to select the appropriate model to represent each offset layer, based on a compromise between goodness of fit to the images and model complexity. More details are given in [12].

The final model obtained using a basic planar model with additional offset primitives is shown in Figure 6. This model is more accurate than the original point-based reconstruction, and contains most of the salient detail in the scene—for instance each of the pillars and windows in the chapel wall are correctly modelled.

Another model reconstructed using the same technique is given in Figure 7. This reconstruction is based on 3 frames from the sequence used in [9]. Although it lacks some of the detail of the model presented in that paper, the castle walls, and some doors and windows, are accurately recovered and compactly represented as a set of planes rather than a dense cloud of points.

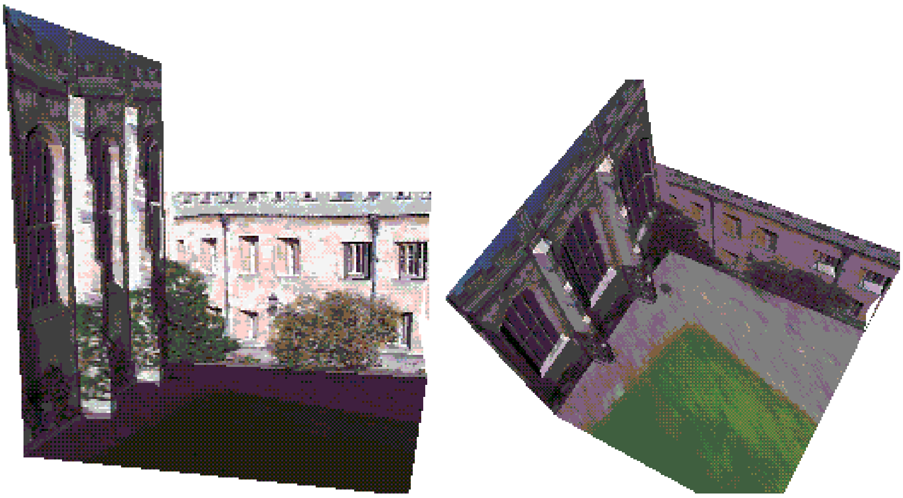


Figure 6: *The refined 3D model. The extruding pillars and inset windows have been accurately modelled by automatically selecting and fitting members from a predefined set of common architectural shapes. The negative log likelihood of this refined model is* 1.80×10^5. *This slight improvement over the model in Figure 4(b) indicates that there is barely enough information in the images to perform this refinement.*

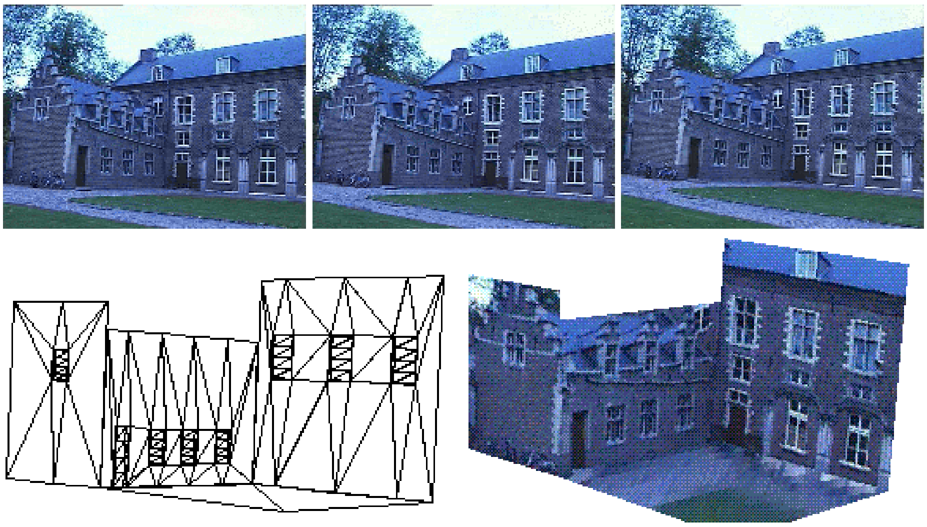


Figure 7: *Frames from the castle sequence (available at* `http://www.esat.kuleuven.ac.be/~pollefey/demos/castle.html`*) used to obtain the model* (top row), *and wireframe and texture mapped views of the 3D model. The castle walls are accurately segmented and reconstructed, and door and window layers are fitted to the walls from a manual initialisation.*

6 Conclusion

This paper has described a technique for generating 3D models of architecture automatically from a small number of images. Knowledge of the nature of architectural structure is used to initialise a plane-based model from one based on 3D points. The plane-based model is optimised using an efficient search algorithm and then refined by the addition of offset layers with constrained shape.

Future work will involve extending this method to handle more complex structure. In particular work will focus on the formulation of more general priors on building shape, including the shape of an overhead view, the shape of each wall and the shape of offset layers within walls.

Acknowledgement

The authors thank Duncan Robertson for the use of his non-linear optimisation software.

References

[1] C. Baillard and A. Zisserman. Automatic reconstruction of piecewise planar models from multiple views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 559–565, June 1999.

[2] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *IEEE Computer Vision and Pattern Recognition*, pages 434–441, 1998.

[3] P.A. Beardsley, P.H.S. Torr, and A. Zisserman. 3d model acquisition from extended image sequences. In *European Conference on Computer Vision*, pages II:683–695, 1996.

[4] R. Cipolla, D. Robertson, and E. Boyer. Photobuilder – 3d models of architectural scenes from uncalibrated images. In *IEEE Int. Conf. on Multimedia Computing and Systems*, 1999.

[5] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24(6):381–395, June 1981.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Alvey Conf.*, pages 189–192, 1987.

[7] I. Jermyn and H. Ishikawa. Globally optimal regions and boundaries. In *ICCV99*, pages 904–910, 1999.

[8] P.R.D.S. Mendonca and R. Cipolla. A simple technique for self-calibration. In *CVPR99*, pages I:500–505, 1999.

[9] M. Pollefeys, R. Koch, M. Vergauwen, and L. van Gool. Metric 3d surface reconstruction from uncalibrated image sequences. In *European Workshop, SMILE*, pages 139–155, 1998.

[10] H.Y. Shum, M. Han, and R. Szeliski. Interactive construction of 3d models from panoramic mosaics. In *IEEE Computer Vision and Pattern Recognition*, pages 427–433, 1998.

[11] C.J. Taylor, P.E. Debevec, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. Technical report, University of California at Berkeley, 1996.

[12] P. Torr, A. Dick, and R. Cipolla. Layer extraction with a bayesian model of shapes. In *ECCV00*, 2000.

[13] P. Torr, R. Szeliski, and P. Anandan. An integrated bayesian approach to layer extraction from image sequences. In *International Conference on Computer Vision*, pages 983–990, 1999.