Automated Architecture Reconstruction from Close-range Photogrammetry*

Tomas Werner, Frederik Schaffalitzky and Andrew Zisserman
Department of Engineering Science
University of Oxford
Oxford, OX1 3PJ, UK
{werner,fsm,az}@robots.ox.ac.uk
http://www.robots.ox.ac.uk/∼vgg

Abstract

We describe a method of automated reconstruction of buildings from a set of uncalibrated photographs. The method proceeds in two steps (i) Recovering the camera corresponding to each photograph and a set of sparse scene features using uncalibrated structure from motion techniques developed in the Computer Vision community. (ii) A novel plane-sweep algorithm which progressively constructs a piecewise planar 3D model of the building. In both steps, the rich geometric constraints present in architectural scenes are utilized. It is also demonstrated that window indentations may be computed automatically.

The methods are illustrated on an image triplet of a college court at Oxford, and on the CIPA reference image set of Zurich City Hall.

Keywords: 3D computer vision, multiple view geometry, plane sweeping, inter image homography

1 Introduction

There has been intensive research effort in the Photogrammetry and Computer Vision communities on reconstruction of architecture from photographs. For example, the following large scale projects dealt with various degrees of automated scene recovery, generally starting from cameras with known calibration and/or range imagery: Ascender [4], Facade [14], IMPACT [1, 6, 10], and RESOLV [13].

In particular, the Facade project demonstrated the high quality of models that could be constructed manually from photographs using a paradigm based on first constructing a polyhedral approximation of the scene and then considered deviations from this approximation. The aim of the work here is an automated Facade modeller – the goal at this stage is first to recover a piecewise planar model that approximates the dominant planes in the scene and their delineation; and then to use these planes to organize the search for perturbations from the plane such as indentations (e.g. windows) and protrusions (e.g. window sills).

We are interested here in architectural scenes which typically contain planes orientated in three dominant directions which are perpendicular to each other, for example the vertical sides of a building and the horizontal ground plane. It is assumed that the scene contains three such principal directions and that the images contain sufficient information to obtain the vanishing points of these directions.

We describe a method which proceeds in two steps: first, the cameras are determined from the images. We assume that the cameras have square pixels and determine the camera matrices using a combination of multiple view matching, and vanishing points corresponding to the three principal scene directions. This is described in section 2.

The second step is to build the piecewise planar model given the cameras. Here we use a "plane sweeping" approach to determine the planes in the principal directions. This plane sweeping strategy builds on previous work on automated reconstruction from aerial views by Baillard *et al* [1]. It is a powerful building block for an architecture reconstruction system enabling the main walls to be recovered efficiently and reliably. Once the main structures are determined, more demanding searches for smaller piecewise planar parts, details in the wall, etc follow. This is described in section 3.

2 Computing the camera matrices

In many photogrammetry applications both the interior and exterior orientation of the cameras are provided. Here we consider the case, which is more commonly the starting point in computer vision applications, where only the images

* This work was supported by EC Project VIBES and an EC Marie Curie fellowship.

(a) (b) (c)

Figure 1: Three images of Merton College, Oxford, acquired with a hand held low cost Olympus digital camera. The images are 1024×768 pixels.

are available and the cameras must be computed directly from these. We need to compute for each view i a 3×4 camera matrix of the form $\mathtt{P}^i = \mathtt{K}^i[\mathtt{R}^i \mid \mathbf{t}^i]$, where $\mathtt{K}$ is an upper triangular 3×3 matrix representing the internal parameters, $\mathtt{R}$ is a rotation matrix and $\mathbf{t}$ a translation. We describe in this section the automatic computation of these cameras from the images.

Determining cameras (without using 3D fiducial points and resectioning) generally proceeds in two stages: first, a projective reconstruction is obtained by determining image points and their correspondences; second, a metric reconstruction is determined from the projective reconstruction using additional constraints.

In more detail, suppose a set of 3D points $\mathbf{X}_j$ is viewed by a set of cameras with matrices $\mathtt{P}^i$. Denote by $\mathbf{x}^i_j$ the coordinates of the j-th point as seen by the i-th camera. Then a projective reconstruction $\{\mathtt{P}^i, \mathbf{X}_j\}$ satisfies $\mathtt{P}^i\mathbf{X}_j \simeq \mathbf{x}^i_j$, and is related to the true scene geometry by an arbitrary 3D projective transformation. A metric reconstruction is related to the true scene geometry by a scaled Euclidean transformation.

In this case determining the projective reconstruction involves obtaining the corresponding image points $\mathbf{x}^i_j$. Determining a metric reconstruction involves obtaining the vanishing points corresponding to the three (orthogonal) principal scene directions. We will illustrate the method for the image triplet shown in figure 1.

2.1 Projective cameras

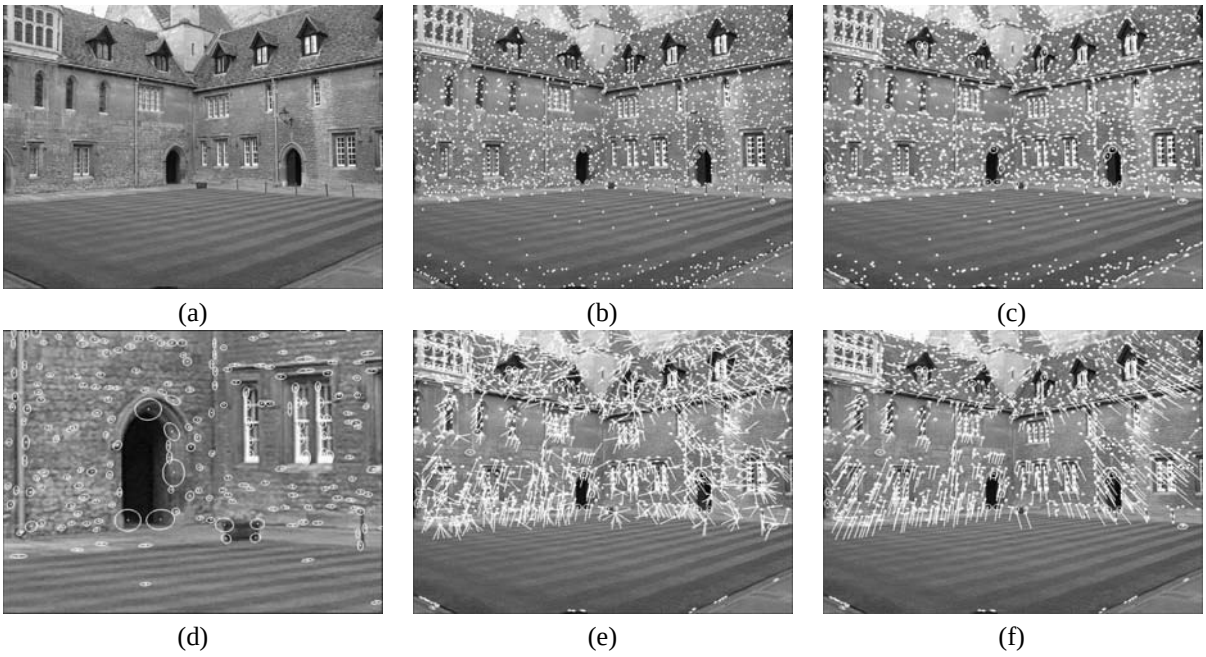
For a significant variety of scene types the epipolar geometry can be computed automatically from two uncalibrated images provided the motion between the views is limited [15, 17]. The computation methods are based on robust statistics and proceed in three steps: first, interest points are detected independently in each image; second, putative point matches are computed between the images based on a measure of proximity and intensity neighbourhood similarity; third, the fundamental matrix (representing the epipolar geometry) and a subset of these matches consistent with the epipolar geometry are determined simultaneously, using a robust estimation algorithm such as RANSAC or LMS. Similarly for limited motion the trifocal geometry can be estimated automatically between image triplets [16]. These methods are reviewed in [8].

We have adapted a number of previous algorithms from the literature, in particular those suited to "wide base line stereo" applications, to be especially suited to architectural scenes. In such scenes there is generally sufficient texture, from bricks, windows etc, to generate a plentiful supply of interest point features; and also sufficient planar patches to enable growing of matches using local planar homographies. The matching algorithm combines ideas from three previous papers: first, detecting interest points at characteristic scales [9]; second, labelling each point by an affine invariant descriptor [2] so that points can be matched on these labels; third, growing matches using planar homographies [11] since surfaces are piecewise planar.

The algorithm is illustrated in figure 2 for the triplet of figure 1. A reconstruction is computed from the resulting matches across the triplet using the method of [12]. The result of this algorithm is a projective reconstruction for the image triplet over 684 points. The bundle adjustment achieves a RMS reprojection accuracy of 0.14 pixels, with a maximum error of 1.01 pixels.

2.2 Metric cameras

The approach used here to upgrade the projective reconstruction to metric is based on determining vanishing points in each image corresponding to the three principal and orthogonal scene directions, and proceeds in three steps. First, the three principal directions (points at infinity) are computed from their images (the vanishing points). This determines the plane at infinity as the plane defined by the three directions, and consequently the projective reconstruction is upgraded to affine. Second, the principal directions are forced to be mutually orthogonal. This determines a reconstruction which differs from metric only by a scaling in each of the principal directions. In the last step, these scalings



(a) (b) (c)

(d) (e) (f)

Figure 2: Steps in computing image point correspondences and projective camera matrices for the triplet of figure 1. (a) The first image of the triplet. (b) Harris corners (2000 per image) and their characteristic scales represented by circles. Note that very few features are found on the grass. (c) The affine invariant neighbourhoods for each corner represented by an ellipse, (d) shows a close up of this. (e) Putative matches to the second image of the triplet based on the invariant descriptor of each corner neighbourhood. A match is indicated by a line from the current position of the corner to the position of the putative match in the other image. A total of 1441 matches are found. (f) The 684 final matches after verification and growing.

are determined (up to an overall scale) using the constraint that the camera pixels are square by linearly minimizing an algebraic residual.

This linear, non-iterative algorithm yields a very good initial estimate of the metric cameras $\mathtt{P}^i = \mathtt{K}^i[\mathtt{R}^i \mid \mathbf{t}^i]$, i.e. both the internal and exterior orientation. For example the computed internal parameters for the first image are: principal point $(580.5, 349.0)$, aspect ratio 1.00055, angle between image axes $89.8944°$, and focal length 1085.4 pixels. Their futher improvement is possible by a bundle adjustment constrained by orthogonality and square pixels assumptions. The typical line sets used to compute the vanishing points are shown in figure 3.

3 Computing the principal scene planes by sweeping

In this section we present a novel strategy for determining planar parts of the scene, typically walls and the ground plane, by "plane sweeping".

(a) (b) (c)

Figure 3: Vanishing point computation. (a), (b) and (c) shows the lines (in white) supporting the vanishing points corresponding to the three principal directions for the thirst image of figure 1. The thick line in (b) and (c) is the horizon computed as the line through the vanishing points.

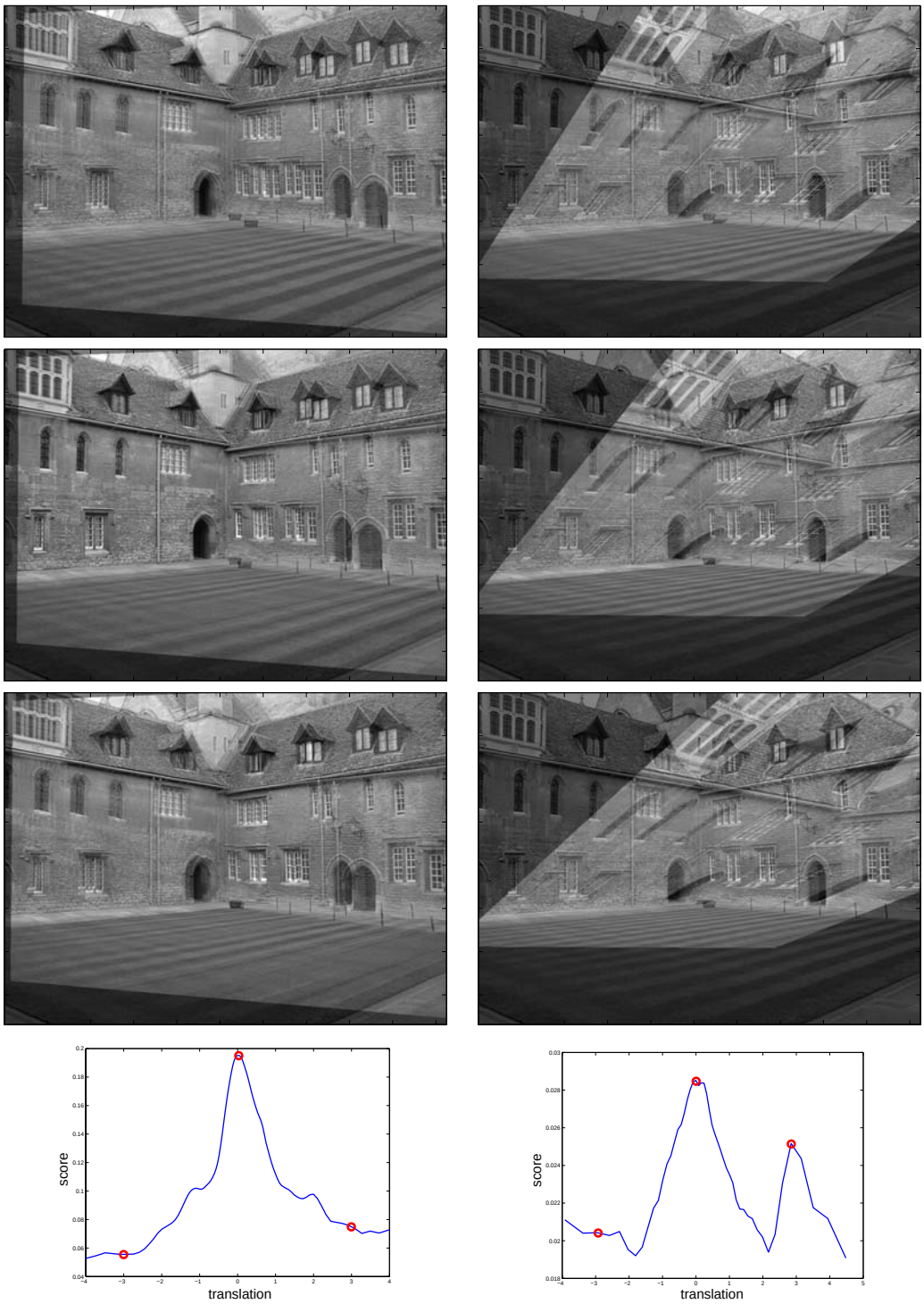


Figure 4: Plane sweeps under translations along the principal scene directions. Each column shows the first and second images of figure 1 superimposed by a homography map corresponding to a translating scene plane. In the first column the translation direction is perpendicular to the left plane, and in the second column the direction is vertical. The circles on the score function at the bottom of each column correspond to the translations ($x-$axis) and image correlations ($y-$axis) of each superposition of the column. In each case the middle translation is the one which best registers the planes, and this is visible because the plane of interest is most "focussed". Note that the repeated structure on the grass gives rise to a second mode in the score function for the ground plane.

Consider the vanishing line of the ground plane. This is the horizon in the image and may be computed from the two vanishing points corresponding to the sets of lines in the scene which are parallel to the ground plane. See figure 3b and c. Given the horizon in two images and the camera matrices the line at infinity which projects to the horizon in each image may be computed. Since all planes which intersect in a common line at infinity are parallel, we can construct a one-parameter family of scene planes parallel to the ground plane, and use these to define a one parameter family of homographies between views. In effect we are translating (sweeping) a scene plane parallel to the ground plane (i.e. vertically) and using this to define the homography map between the images.

When the swept plane corresponds to the actual ground plane, then the intensity function in one image and the intensity function from another image transformed by this homography will be highly correlated in the region corresponding to the ground. By determining the swept plane which produces the highest correlation (measured by an appropriate robust function described below) the true position of the ground plane is determined. The sweeping is illustrated for two of the principal directions in figure 4.

The cross-correlation is measured as follows: first image points at significant image gradients are determined; second, a sample of these points are selected as the key points at which to measure the cross-correlation. Only these points are used in order to improve the signal to noise ratio by discarding uninformative points in homogeneous image regions. Approximately 10% of image pixels were selected as the keypoints. At each of the key points the cross-correlation is computed between the point's neighbourhood and the corresponding neighbourhood in the other image mapped by the homography induced by the sweeping plane. The cross-correlation is computed between all pairs of images, and the result averaged. A score function consists of the cross-correlations averaged over all points as a function of plane translation. Typical score functions are shown in figure 4.

The idea of using a point-to-point map to correctly measure the cross-correlation is not new – indeed Gruen proposed this in the Photogrammetry literature in 1985 [7]. However, the map in that case was an affinity (an approximation), and its computation required a search over *six* parameters. Here no approximation is involved, and the homography (in general specified by *eight* parameters) is computed by a one-parameter search. Collins [3] used a homography in this manner to search for line matches above a ground plane by sweeping a plane in the vertical direction. However, the exterior orientation of the cameras was known. A similar method was employed in [5] to search for lines on vertical facades.

The most significant result here is determining the plane corresponding to the ground plane, since very few features are detected on this plane. Consequently it is very difficult to determine the ground plane from feature matches alone.

Given these three (infinite) planes it is then possible to determine the partial delineation of each plane from their intersection (figure 7a); and hence to texture map the appropriate parts of the images onto a three dimensional model (figure 7b,c).

4 Determining perturbations from the principal planes

Having computed the principal planes and their partial delineations, we are now able to concentrate on particular image regions (and thereby improve the signal to noise), and use the delineated scene planes to organize the search for model refinements. In particular it will be shown here that rectangular window indentations can be modelled automatically.

The idea is to determine regions of the plane which do not coincide with the coarsely fitted scene plane, and then to model these regions as rectangles aligned with the principal scene directions. Note, each region is modelled independently as opposed to assuming that all windows are at the same depth and have equal size and are arranged on a regular grid.

Points which lie behind the fitted plane (indentations) are determined by thresholding depths of individual keypoints. The threshold value is obtained from the score function recomputed for the image region corresponding to the current facade, see figure 5a. Two modes are clearly discernable in the function – one corresponding to the coarse facade plane, and the other to the window plane.

The keypoints labeled by thresholding as being behind the wall are shown in figure 5d. Contiguous regions (corresponding to each of the windows and doors) are then computed by robustly clustering these points. The fitted plane has essentially simplified this task to that of clustering a set of pixels in a 2D image as the analysis can be carried out on a rectified version of the image (where the principal scene directions are orthogonal in the image). Standard image processing methods, such as simple operations of binary morphology, are used. Rectangular boxes are then fitted to the clusters. The resulting windows boundaries are shown in figure 5e.

The window boundaries are then refined using a further correlation based search, but now concentrated in the vicinity of the putative window boundary. For each boundary a one-dimensional search is carried out perpendicular to that boundary. For example for a vertical boundary the search is horizontal. Two score functions are computed, one based on the homography induced by the wall plane, the other based on the homography induced by the window plane, as illustrated in figure 6. For the wall plane the score is high when the pixels in the rectangle belong to the wall and small otherwise. Conversely, for the window plane homography the similarity score is high for window pixels,

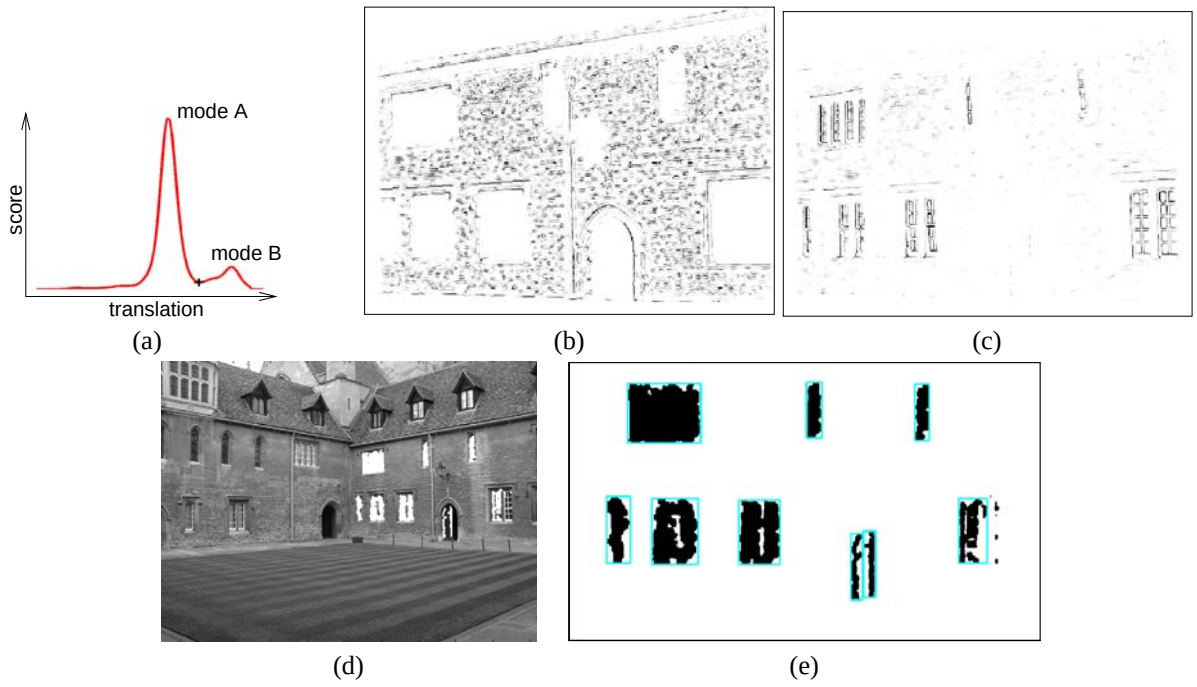


Figure 5: (a) Aggregated similarity score in the neighbourhood of the wall plane computed over three views. Mode A corresponds to the wall plane, and B to the plane of the windows. The magnitude of the similarity scores for individual keypoints are represented by pixel darkness in (b) for the swept plane at the position of mode A, and (c) for the position of mode B. The keypoints belonging mostly to the wall plane score highly (are dark) in (b) and those belonging to the window plane score highly in (c). Points belonging to windows, (d), are obtained by thresholding depths at the value denoted by the cross in plot (a). (e) Rectangles robustly fitted to clustered points shown on the rectified facade.

and low for wall ones. The product of these two similarity scores peaks at the actual window boundary, as shown in the figure.

The facade model is then updated with these rectangular indentations, and texture mapped from the appropriate images. The resulting piecewise planar model is shown in figure 7d,e,f.

Figure 8 shows four views of one facade of the Zurich City Hall from the CIPA image set. The model with window indentations computed using the set of sweeping algorithms described above is shown in figure 9.

5 Discussion

This paper has sketched the idea of using lines at infinity for plane sweeping. This adds another method, to those already available, for determining planes in a structured scene. Its advantage over other existing methods (such as robust plane fitting to point clouds) is that feature correspondences are not required.

References

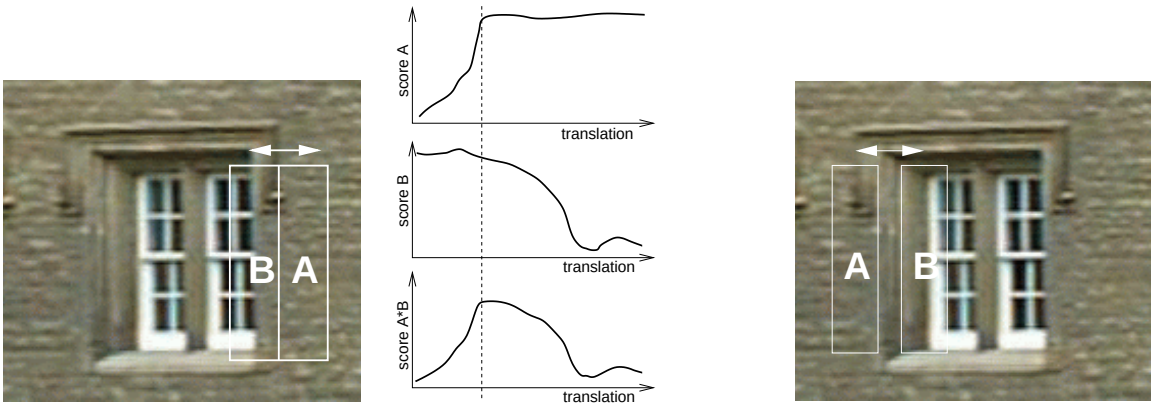
[1] C. Baillard, C. Schmid, A. Zisserman, and A. Fitzgibbon. Automatic line matching and 3D reconstruction of buildings from multiple views. In *ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery, IAPRS Vol.32, Part 3-2W5*, pages 69–80, September 1999.

[2] A. Baumberg. Reliable feature matching across widely separated views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 774–781, 2000.

[3] R. T. Collins. A space-sweep approach to true multi-image matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 358–363, 1996.

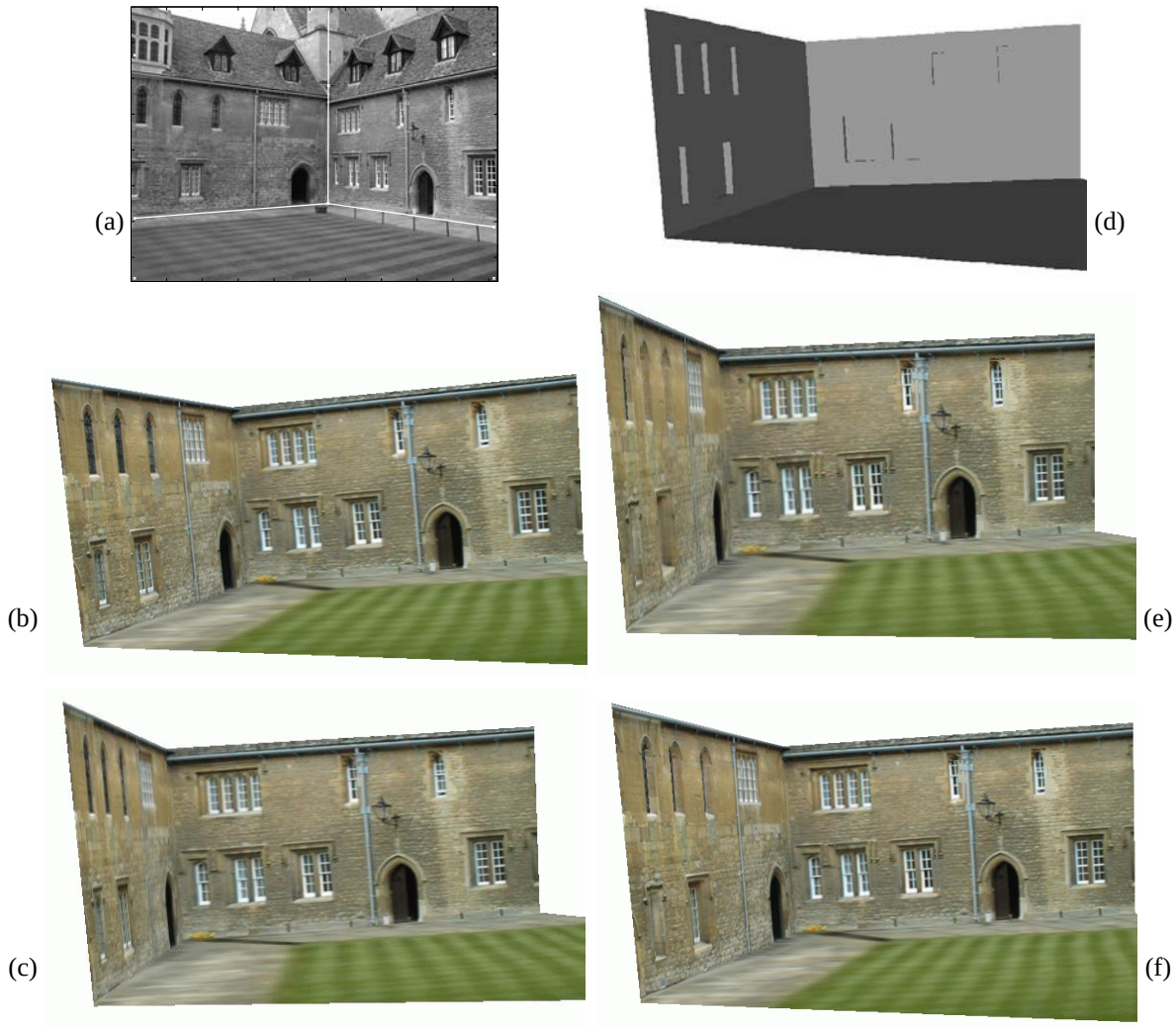
[4] R. T. Collins, C.O Jaynes, , Y-Q Cheng, X. Wang, F. Stolle, E. M. Riseman, and A. R. Hanson. The ascender system: Automated site modeling from multiple images. *Computer Vision and Image Understanding*, 72(2):143–162, 1998.

[5] S. Coorg and S. Teller. Extracting textured vertical facades from controlled close-range imagery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado*, pages 625–632, 1999.



(a) (b) (c)

Figure 6: Refining the position of the vertical window boundaries. (a) Search regions consisting of two adjacent rectangles which are translated horizontally. In rectangle A the similarity score is based on the wall plane. In rectangle B the window plane is used. (b) plots the two scores and their product, which peaks at the actual window boundary location (the dashed line). (c) For non-occluding window edges (the left one in this case), a gap of appropriate width is inserted between the rectangles.



(a) (d) (b) (e) (c) (f)

Figure 7: (a) The lines of intersection of the three computed planes projected onto the third image of figure 1. (b,c) Two views of the texture mapped coarse 3D VRML model, consisting of wall planes and the ground plane, partially delineated. The automatically computed 3D model augmented with indented windows is shown in (d) untextured, and texture mapped from the appropriate regions of the images in (e) and (f). Observe the difference between windows not modelled (b,c) and modelled (e,f) as the viewpoint changes.

Figure 8: Four views of one side of the Zurich City hall which are used in the plane sweep to determine the window indentations.

(a) (b) (c)

Figure 9: The shaded (a) and textured (b,c) model of the Zurich City hall. Observe the effect of window parallax in (b,c).

[6] M. Fradkin, M. Roux, and H. Maître. Building detection from multiple views. In *ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery*, September 1999.

[7] A.W. Gruen. Adaptative least squares correlation: a powerful image matching technique. *S. Afr. Journal of Photogrammetry, Remote Sensing and Cartography*, 3(14):175–187, 1985.

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[9] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. 8th International Conference on Computer Vision, Vancouver, Canada*, 2001.

[10] T. Moons, D. Frère, J. Vandekerckhove, and L. Van Gool. Automatic modelling and 3D reconstruction of urban house roofs from high resolution aerial imagery. In *Proc. 5th European Conference on Computer Vision, Freiburg, Germany*, pages 410–425, 1998.

[11] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proc. 6th International Conference on Computer Vision, Bombay, India*, pages 754–760, January 1998.

[12] F. Schaffalitzky, A. Zisserman, Hartley, R. I., and P. H. S. Torr. A six point solution for structure and motion. In *Proc. European Conference on Computer Vision*, pages 632–648. Springer-Verlag, June 2000.

[13] V. Sequeira, K.C. Ng, E. Wolfart, J.G.M. Goncalves, and D.C. Hogg. Automated reconstruction of 3D models from real environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, (54):1–22, 1999.

[14] C. Taylor, P. Debevec, and J. Malik. Reconstructing polyhedral models of architectural scenes from photographs. In *Proc. 4th European Conference on Computer Vision, Cambridge*. Springer-Verlag, 1996.

[15] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24(3):271–300, 1997.

[16] P. H. S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.

[17] Z. Zhang, R. Deriche, O. D. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78:87–119, 1995.