Automatic reconstruction of piecewise planar models from multiple views

C. Baillard and A. Zisserman
Dept. of Engineering Science, University of Oxford,
Oxford OX13PJ, England
{caroline,az}@robots.ox.ac.uk

Abstract

A new method is described for automatically reconstructing 3D planar faces from multiple images of a scene. The novelty of the approach lies in the use of inter-image homographies to validate and best estimate the plane, and in the minimal initialization requirements — only a single 3D line with a textured neighbourhood is required to generate a plane hypothesis. The planar facets enable line grouping and also the construction of parts of the wireframe which were missed due to the inevitable shortcomings of feature detection and matching.

The method allows a piecewise planar model of a scene to be built completely automatically, with no user intervention at any stage, given only the images and camera projection matrices as input. The robustness and reliability of the method are illustrated on several examples, from both aerial and interior views.

1. Introduction

The automatic reconstruction of 3D planar faces from multiple views has a wide range of applications, from exterior views of urban scenes to indoor environment analysis to piecewise planar objects. The target application of this paper is the 3D reconstruction of roofs from aerial images of urban areas, but the method is not restricted to this case.

Reconstruction of buildings from aerial images has received continual attention in the photogrammetry and computer vision literature. One approach is to compute a dense digital elevation model from multiple images using correlation based stereo. However, the resulting dense depth map is generally insufficiently accurate or complete to enable the precise shape of buildings to be recovered [1, 2, 5]. Thus most approaches have focused on reconstruction of specific building models: rectilinear shapes [4, 8, 10, 11], flat roofs [2], or parametric models [6, 13]. Recently, more generic reconstruction approaches involving multiple high-resolution images have been proposed [3, 9]. The difficulty of reconstruction in urban environments comes from the complexity of the scene: the buildings are dense and varied, and the resulting image boundaries often have poor contrast. Consequently, feature detectors fragment or miss boundary lines, and only an incomplete 3D wireframe can be obtained.

The key idea here is to determine 3D planar facets by using both features (lines) and their image neighbourhoods over multiple views. These surface facets then enable both line grouping and, by plane intersection, the creation of lines which were missed during feature detection. The particular novelty of the approach is in the use of inter-image homographies (plane projective transformations) to robustly estimate the planar facets. The approach requires minimal image information since a plane is generated from only a line correspondence and its image neighbourhood. In particular two lines are not required to instantiate a plane. These minimal requirements and avoidance of specific building models facilitate the automatic reconstruction of objects with quite subtle geometry located within a complex environment.

The problems caused by missing features in piecewise planar reconstruction are illustrated by the images of figure 1 and the detail in figure 6a. The correct roof model in this case is a four plane “hip” roof [13]. However, the oblique roof ridges are almost invisible in any view, and certainly are not reliably detected by an edge or bar detector with only local neighbourhood support. Consequently, ‘classical’ plane reconstruction algorithms which proceed from a grouping of two or more coplanar 3D lines [3, 9], will produce a flat roof, or at best a two plane “gable” roof if the central horizontal ridge edge is detected — however the two smaller faces will be missed.

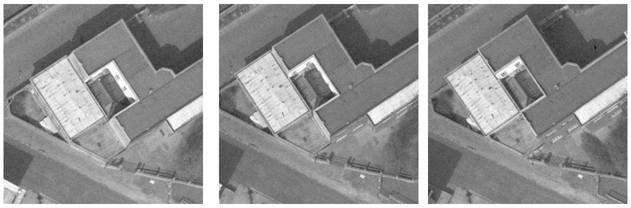


Figure 1. Three overlapping aerial views. The images are 600×600 pixels, one pixel corresponding to a ground length of 8.5cm. The camera moves in a line at a height of 1300m, with about 300m between successive views.

1.1. Overview of the method

Input data. The method will be illustrated on the quasi-nadir images shown in figure 1 (3 of the 6 images are shown). From these images only roof planes of buildings can be extracted since vertical walls are generally not visible. The camera projection matrix is known for each view. In this case the projection matrices are metric calibrated. However, the method requires only projective information.

Line matching. The 2D image lines are obtained by applying a local implementation of the Canny edge detector (with subpixel accuracy), detecting tangent discontinuities in the edgel chains, and finally straight line estimation by orthogonal regression.

Then lines in 3D are generated by using an implementation of the line matching algorithm for 3 views described in [12]. Matches are disambiguated by a geometric constraint over 3 views (using epipolar geometry and trifocal geometry), together with a photometric constraint based on line intensity neighbourhoods. Here the line matching is extended to six views. Figure 2 shows the result of the line matching on the data set of figure 1. Note that some of the scene lines are missing, and some of the recovered lines are fragmented.

Producing piecewise planar models. There are three main stages, which will be illustrated on the building of figure 6a:

1. *Computing reliable half-planes* defined by one 3D line and similarity scores computed over all the views (section 2). This is the most important and novel stage of the algorithm.

2. *Line grouping and completion* based on the computed half-planes (section 3). This involves grouping neighbouring 3D lines belonging to the same half-plane, and also creating new lines by plane intersection.

3. *Plane delineation and verification* where the lines of the previous stage are used to delineate the plane boundaries (section 4).

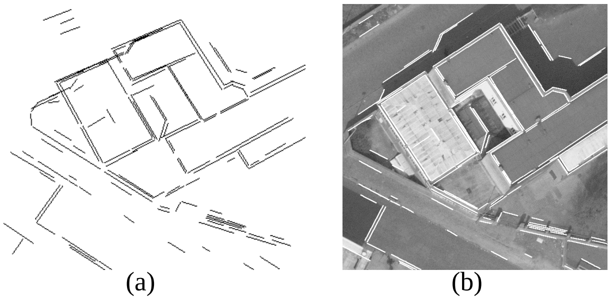


Figure 2. Line matching. (a) 137 lines are matched automatically over 6 views. Their 3D position (shown) is determined by minimizing reprojection error over each view in which the line appears. (b) The lines projected onto the first image of figure 1.

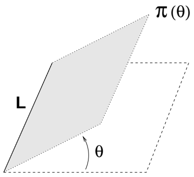


Figure 3. The one-parameter family of half-planes containing the 3D line $\mathtt{L}$. The family induces a one-parameter family of homographies between any pair of images.

2. Computing half-planes

Principles and objectives. Given a 3D line, there is a one-parameter family of planes $\boldsymbol{\pi}(\theta)$ containing the line (see figure 3). As each plane defines a (planar) homography between two images, the family also defines a one-parameter family of homographies $\mathtt{H}(\theta)$ between any pair of images. Each side of the line can be associated with a different half-plane.

Our objective is therefore to determine for each line side whether there is an attached half-plane or not, and if there is we want to compute a best estimate of θ. We wish to employ only the minimal information of a single 3D line and its image neighbourhood. Essentially we are hypothesising a planar facet attached to the line, and verifying or refuting this model hypothesis using image support over multiple views.

Method. The existence of an attached half-plane and a best estimate of its angle is determined by measuring image similarity over multiple views. The geometry is illustrated in figure 4. Given θ, the plane $\boldsymbol{\pi}(\theta)$ defines a point to point map between the images. If the plane is correct then the intensities at corresponding pixels will be highly correlated.

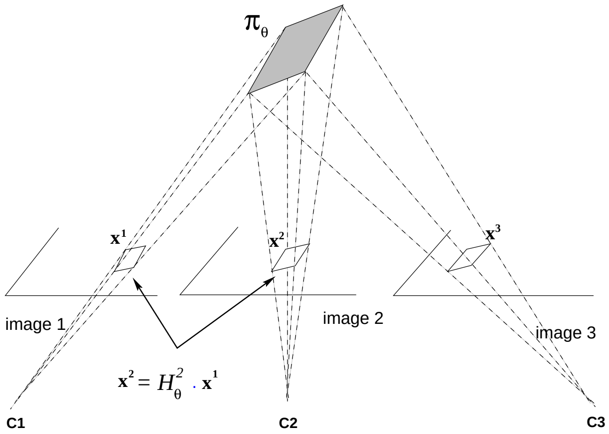


Figure 4. Geometric correspondence between views. Given θ, the homography $\mathtt{H}^i(\theta)$ determines the geometric map between a point in the first image and its corresponding point in image i.

In more detail, the plane $\boldsymbol{\pi}(\theta)$ attached to a 3D line $\mathbf{L}$ is assessed by a similarity score computed over the 6 images according to the homographies defined by θ. Given the plane $\boldsymbol{\pi}(\theta)$ there is a homography represented by 3×3 matrix $\mathtt{H}^i(\theta)$ between the first and ith view, so that corresponding points are mapped as $\mathbf{x}^i = \mathtt{H}^i\mathbf{x}$, where $\mathbf{x}$ and $\mathbf{x}^i$ are image points represented by homogeneous 3-vectors. The homography matrix is obtained from the 3×4 camera projection matrices for each view. For example, if the projection matrices for the the first and ith views are $\mathtt{P} = [\mathtt{I} \mid \mathbf{0}]$ and $\mathtt{P}^i = [\mathtt{A}^i \mid \mathbf{a}^i]$ respectively, and 3D points $\mathbf{X}$ on the plane satisfy $\boldsymbol{\pi}^\top \mathbf{X} = 0$, where the plane is represented as a homogeneous 4-vector in the world frame, then [7]

$$\mathtt{H}^i = \mathtt{A}^i + \mathbf{a}^i\mathbf{v}^\top \text{ where } \mathbf{v} = -\frac{1}{\pi_4}(\pi_1, \pi_2, \pi_3)^\top$$

provided $\pi_4 \neq 0$. Note, $\mathbf{v}$ is independent of the view i.

The similarity score function has been designed to be selective, and also robust to occluded portions and irrelevant points. It is defined as:

$$Sim(\theta) = \sum_{\text{view } i \text{ valid}} \int_{POI} w(\mathbf{x}) Cor^2(\mathbf{x}, \mathtt{H}^i(\theta)\mathbf{x})$$

and ranges between $(0, 1)$. Figure 5 shows two typical examples of score functions. In the following, the various terms and parameters are described and motivated.

First, it is necessary to determine a texture point set in order to produce a selective and discriminating similarity function of θ. Consider the case that the intensity of the image is locally homogeneous, then correlation between images is similar for any θ and provides no discrimination. However, at locally textured regions this problem will not arise. Correlation is thus computed only in the neighbourhood of textured Points Of Interest (POI). These points are selected by applying an edge detector to the first image, with a very low threshold on gradient (an example is given in figure 6). The edges are then linked and regularly sampled. The choice of the first view is arbitrary and can be automated by selecting the most textured image.

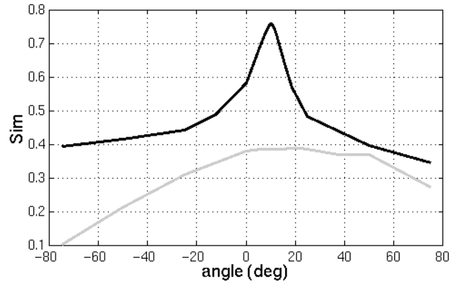
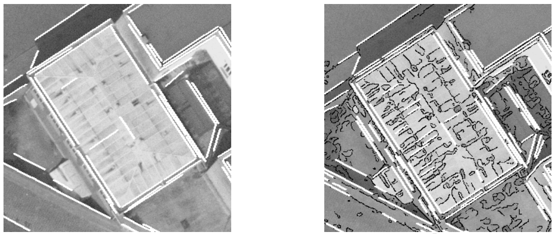


Figure 5. Example of similarity score functions $Sim(\theta)$. The black curve corresponds to a valid plane, whereas the grey one is rejected. The following validity criteria are used: maximum value of the function $Sim(\theta^{max}) \geq 0.4$, absolute value of the estimated second derivative around the maximum $|Sim''(\theta^{max})| \geq 4.0$, and global amplitude $Sim(\theta^{max}) - Sim(\theta^{min}) \geq 0.2$.



(a) (b)

Figure 6. Points of interest. (a) Detail of figure 2 with projected 3D lines (white). This building is used to illustrate the reconstruction method. The correct reconstruction is a four plane hip roof. (b) Detected edges (black) after applying an edge detector with a very low threshold on gradient. These edges provide the points of interest.

The correlation term $Cor(\mathbf{x}, \mathbf{x}^i)$ is the centred normalised cross-correlation between $\mathbf{x}$ in the first view and $\mathbf{x}^i$ in the ith view, evaluated over the points of interest. Cross-correlation is used because empirically it is highly selective on θ over textured intensity regions. The correlation is squared in order to give more weight to high scores, and therefore to be less sensitive to low scores which can locally occur in some views (in occluded areas for instance).

The weighting factor $w(\mathbf{x})$ is inversely proportional to the distance of the point $\mathbf{x}$ from the line $\mathbf{L}$ projected onto the

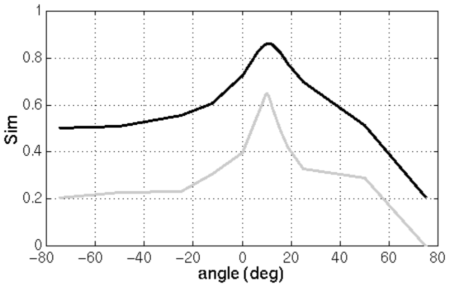


Figure 7. Effect of the base line on 2-view similarity scores. The black curve corresponds to a short baseline between views, the grey curve to a wide one (same half-plane). A short baseline leads to high maxima (low distorsion between the images), but often located with a poor accuracy (wide peak); in contrast, a wide baseline is more likely to produce accurate maxima, but with a lower score. However, the maxima generally differ by less than 5^o.

first view. This weighting provides some robustness, since it gives more weight to points which are closer to the line, and consequently are less likely to belong to other planes. Finally, additional robustness is provided by only including *valid* views in the summation. Valid views are those which have at least a threshold number of high correlation scores at points of interest, thereby rejecting views where the plane might be occluded. Averaging the scores over views exploits the complementarity of the short and wide baseline separations (see figure 7).

The optimal angle θ is computed by searching for the maximum of the function $Sim(\theta)$ over a range $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$. The algorithm used is recursive sub-division with a termination criterion of $\Delta\theta < 1^o$. The half-plane hypothesis is accepted or rejected as valid according to the characteristics of $Sim(\theta)$ as shown in figure 5. The line side is then classified as supporting or not supporting a half-plane. For example, an occluding edge would not have a half-plane attached on the occluded side.

Results of half-plane detection. Figure 8 shows all the half-planes which are hypothesised on the example building. All parts of the roof of the main building are detected, whereas no valid planes are detected for the walls within the considered angle interval (we are not aiming to reconstruct vertical walls). Occasionally erroneous half-planes arise at shadows, but these are removed in the subsequent stages.

3. Grouping and completion of 3D lines based on half-planes

The computed half-planes are now used to support line grouping and the creation of new lines.

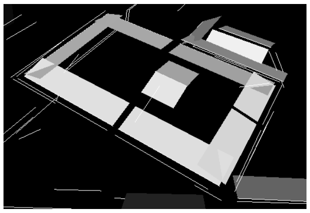


Figure 8. Detected half-planes over the interval $[-75^o; +75^o]$.

Collinear grouping (figure 9). Two collinear lines which have attached coplanar half-planes are merged together. The result of the collinear grouping of half-planes of figure 8 is shown in figure 11a.

Coplanar line and half-plane grouping (figure 10). Any line which is neighbouring and coplanar with the current plane is associated with it (see the example of figure 11b).

Creating new lines by plane intersections (figure 12). New lines are created when two neighbouring planes intersect in a consistent way. This is very important as it provides a mechanism for generating additional lines which may have been missed during image feature detection (see the example of figure 13).

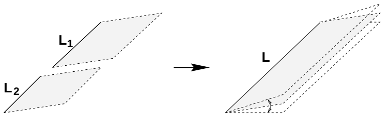


Figure 9. Collinear grouping. The optimal plane angle is recomputed for the merged line, again using $Sim(\theta)$ as described in section 2. This is more accurate than, for instance, averaging angles.

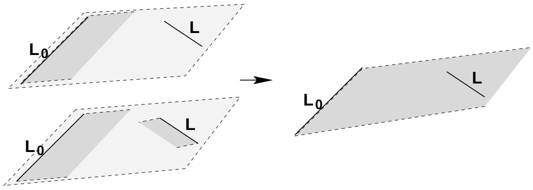


Figure 10. Coplanar line and half-plane grouping. In the top case, $\mathbf{L}$ belongs to the half-plane $\boldsymbol{\pi}(\mathbf{L}_0)$, and a new plane is computed by orthogonal regression to a regular point sampling of $\mathbf{L}_0$ and $\mathbf{L}$. In the bottom case, $\mathbf{L}$ has an attached but consistent half-plane, therefore the two plane hypotheses are merged into a new unique plane, also computed by orthogonal regression.

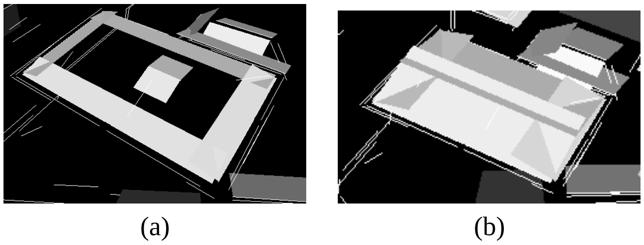


Figure 11. Line grouping. (a) Collinear grouping reduces the 9 planes prior to grouping to only 6. (b) Coplanar grouping and plane merging reduces the number of planes further so that only 4 remain. These are the correct four planes which define the roof, but at this stage the plane boundaries are not delineated.

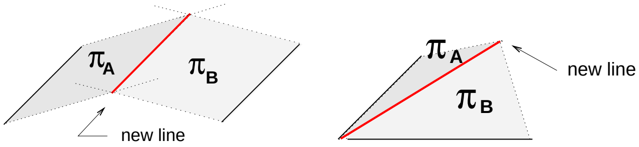


Figure 12. Creation of new lines when two planes intersect.

4. Plane delineation and verification

In order to produce a piecewise planar model of the scene a closed delineation is required for each plane. For this purpose, it is necessary to determine its *border* (or bounding) *lines*. The initial support line of a plane is a natural border line. Additional border lines are created as shown in figure 14. A closed delineation is then computed by using heuristic grouping rules [9, 10, 13] to associate border lines. If only one border line has been detected, then the plane is rejected, and this provides a very efficient culling mechanism for removing erroneous half-planes. Figure 15 shows both the 2D delineation and a 3D view of the roof produced for the building of figure 6.

Each delineated 3D face so produced is then verified by assessing intensity similarity over the complete image set, at corresponding points within the projected delineation. This verification step removes fallacious planes, for example those which erroneously bridge two buildings.

Finally, occlusion reasoning is used to signal conflicts between possibly inconsistent plane hypotheses. A conflict occurs between two hypotheses when their projections onto an image significantly overlap, i.e. when one of them is occluded by the other. Where conflicts arise, they are resolved based on a confidence score determined from the ratio of the length of detected 3D lines to hypothesised lines in the delineation.

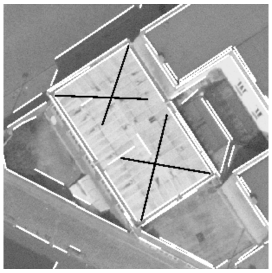


Figure 13. Example of new lines. The black lines are created by plane intersection.

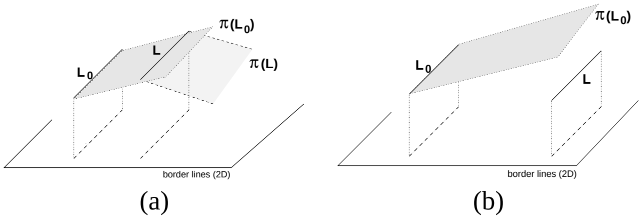


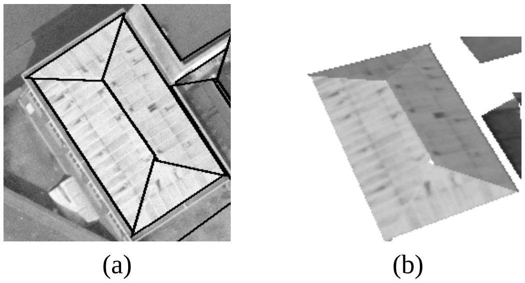
Figure 14. Plane delineation: border line computation. (a) The line $\mathbf{L}$ lies in the plane $\boldsymbol{\pi}(\mathbf{L}_0)$ but has an attached plane which is not consistent with it, therefore it is stored as a border line; (b) The line $\mathbf{L}$ does not belongs to the plane $\boldsymbol{\pi}(\mathbf{L}_0)$ but it is stored as a border line.

5. Results

Figure 16 shows the 3D reconstruction of the full scene of figure 1. Figure 17 shows results on a much larger example. Note that complicated and unusual roofs (for example the factory in the upper part of the figure) have been completely recovered. This also demonstrates how little radiometric texture is required by the method, since roofs with virtually homogeneous intensity are retrieved. Only two roofs are missed in the entire scene.

The versatility of the method is demonstrated on figure 18 where the images are of an indoor scene, at a different scale and under differing photometric conditions to those of the aerial images. The main planes of the scene are correctly retrieved.

Performance. The quality of the reconstruction is governed by the accuracy of the input line set. If lines are missing in the input set, then a proportion of these will be generated by plane intersection. If there are erroneous lines in the input set, then many of these will be culled, with their associated half-planes, in the final verification stages. Consequently, the method is robust to a number of missing and erroneous lines. However, in general superior performance is achieved if too many, rather than too few, lines are supplied. This



(a) (b)

Figure 15. Example of reconstructed roof. (a) Delineation of the validated roofs projected onto the first image; (b) 3D view with texture mapping.

is because a line is the only mechanism for instantiating a plane hypothesis, and if lines are missing then entire planes may be missed.

Of the three stages of the method, the half-plane detection stage is the most robust and is also the most expensive. This stage requires very few parameters to be specified. When a face is well textured (as in the case of the example building roof of figure 6), the angle of the initial half-plane is estimated to an accuracy of better than 2^o. When there is little texture, the accuracy can decrease to 5^o, but a higher accuracy is determined during the coplanar grouping stage. The computation time can be reduced by using a sub-sampling of the points of interest, as well as an improved optimization process (based for example on Newton's method rather than recursive subdivision).

In grouping operations, thresholds on distances are avoided by defining a topological neighbourhood between projected lines, which also enables quick access to neighbours. The plane delineation stage is the least robust to changes in scene type because it involves heuristic grouping rules.

6. Future work and extensions

The results demonstrate that it is possible to automatically reconstruct piecewise planar scenes from multiple images using quite minimal information. We are currently investigating the following areas:

First, the similarity score, although successful, is empirically based. An alternative approach would be to develop a probabilistic framework and compute a posterior estimate for the plane angle and the likelihood that a plane is there.

Second, the approach of this paper is complementary to roof generation systems based on the grouping of at least two neighbouring coplanar lines [3, 9]. However it is not an alternative to such systems, and an efficient and robust approach would use both. This requires an architecture for a cooperating strategy to be developed.

Third, parametrized models (for example a hip roof) should be instantiated from the reconstructed roofs whenever possible. The advantage is that such models can be optimally fitted to the images by (robustly) minimizing reprojection error over all views. This should improve the accuracy of the plane delineation.

Acknowledgements. We are very grateful to Dr A. Fitzgibbon for his assistance with software and helpful discussions. Financial support was provided by EU Esprit Project Impact.

References

[1] C. Baillard, O. Dissard, and H. Maître. Segmentation of urban scenes from aerial stereo imagery. In *Proc. ICPR*, pages 1405–1407, Aug 1998.

[2] M. Berthod, L. Gabet, G. Giraudon, and J. L. Lotti. High-resolution stereo for the detection of buildings. In A.Grün, O.Kübler, and P.Agouris, editors, *Automatic Extraction of Man-Made Objects from Aerial and Space Images*, pages 135–144. Birkhäuser, 1995.

[3] F. Bignone, O. Henricsson, P. Fua, and M. Stricker. Automatic extraction of generic house roofs from high resolution aerial imagery. In *Proc. ECCV*, pages 85–96, 1996.

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

[5] S. Girard, P. Guérin, H. Maître, and M. Roux. Building detection from high resolution colour images. In *Int. symp. on Remote Sensing, Barcelona*, 1998.

[6] N. Haala and M. Hahn. Data fusion for the detection and reconstruction of buildings. In *Automatic Extraction of Man-Made Objects from Aerial and Space Images*, pages 211–220. Birkhäuser, 1995.

[7] Q. T. Luong and T. Vieville. Canonical representations for the geometries of multiple projective views. *CVIU*, 64(2):193–229, September 1996.

[8] J. McGlone and J. Shufelt. Projective and object space geometry for monocular building extraction. In *Proc. CVPR*, pages 54–61, 1994.

[9] 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. ECCV*, pages 410–425, 1998.

[10] S. Noronha and R. Nevatia. Detection and description of buildings from multiple images. In *Proc. CVPR*, pages 588–594, 1997.

[11] M. Roux and D. M. McKeown. Feature matching for building extraction from multiple views. In *Proc. CVPR*, 1994.

[12] C. Schmid and A. Zisserman. Automatic line matching across views. In *Proc. CVPR*, pages 666–671, 1997.

[13] U. Weidner and W. Förstner. Towards automatic building extraction from high-resolution digital elevation models. *ISPRS j. of Photogrammetry and Remote Sensing*, 50(4):38–49, Aug 1995.

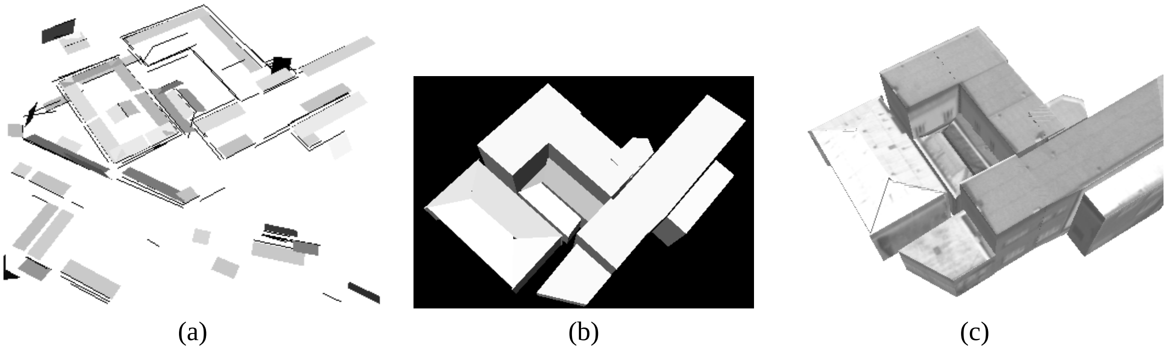


Figure 16. Results on the full example scene. (a) 49 detected half-planes from 137 3Dlines (b) Delineation of the final roofs projected onto the first image; (c) 3D model of the scene, with texture mapping (12 roof planes). The vertical walls are produced by extruding the roof's borders to the ground plane.

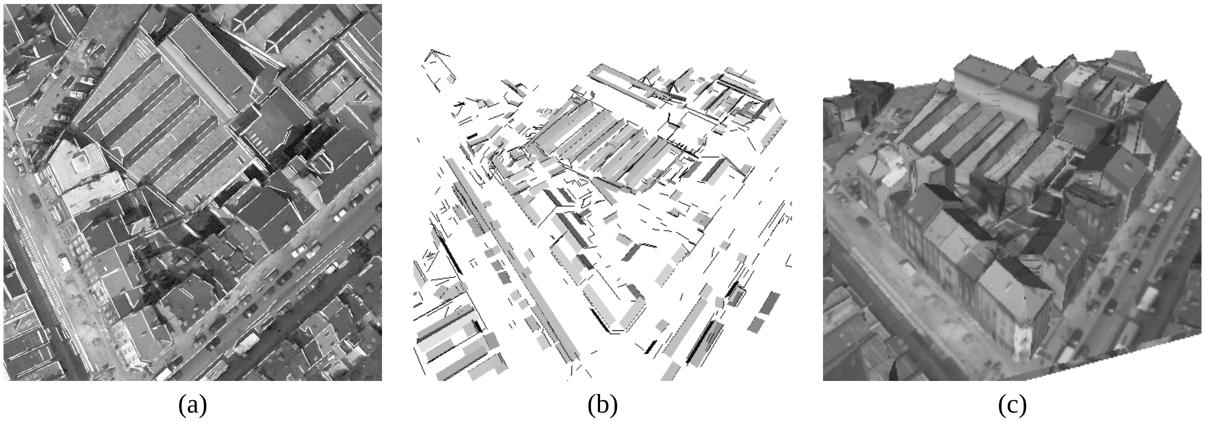


Figure 17. Results on a large aerial scene.(a) One of the 6 images (size 1200 × 1200 pixels) and the 739 projected 3D lines. (b) Detected half-planes (267). (c) 3D model of the scene, with texture mapping (180 roof planes).

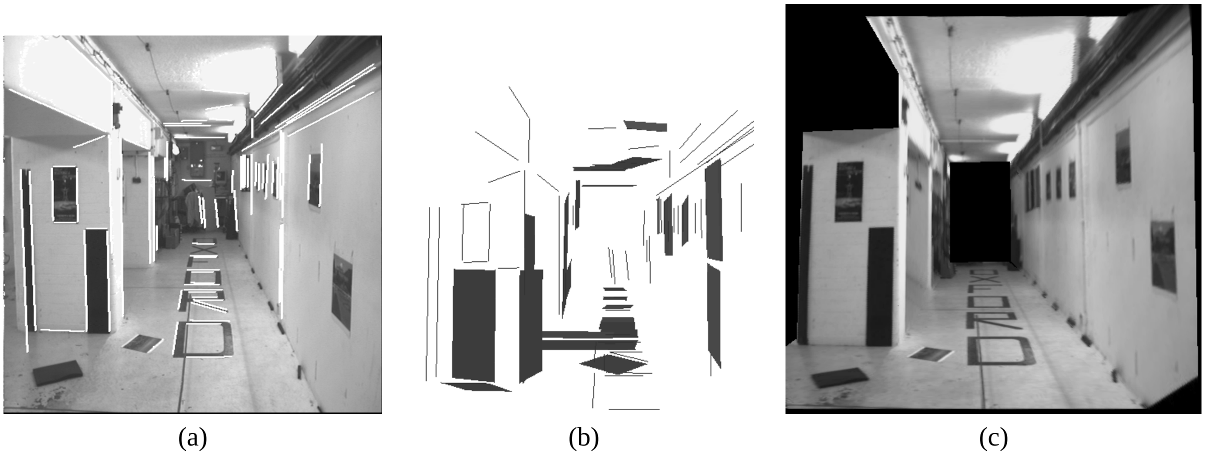


Figure 18. Results on an indoor scene.(a) One of the 6 images (size 512 × 512 pixels) and the 87 projected 3D lines. (b) Detected half-planes (30). (c) 3D model of the scene, with texture mapping (5 planes).