AUTOMATIC LINE MATCHING AND 3D RECONSTRUCTION OF BUILDINGS FROM MULTIPLE VIEWS

C. Baillard, C. Schmid, A. Zisserman and A. Fitzgibbon
Dept. of Engineering Science, University of Oxford,
Oxford OX13PJ, England
Ph.: +33-1865-273127, Fax: +33-1865-273908
E-mail: {caroline,az,awf}@robots.ox.ac.uk, Cordelia.Schmid@inrialpes.fr

KEY WORDS: line matching, building reconstruction, urban areas, multiple views, planar homography.

ABSTRACT

This paper describes two developments in the automatic reconstruction of buildings from aerial images. The first is an algorithm for automatically matching line segments over multiple images. The algorithm employs geometric constraints based on the multi-view geometry together with photometric constraints derived from the line neighbourhood, and achieves a performance of better than 95% correct matches over three views. The second development is a method for automatically computing a piecewise planar reconstruction based on the matched lines. The novelty here is that a planar facet hypothesis can be generated from a single 3D line, using an inter-image homography applied to the line neighbourhood. The algorithm has successfully generated near complete roof reconstructions from multiple images. This work has been carried out as part of the EC IMPACT project. A summary of the project is included.

1 INTRODUCTION

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 (Berthod et al., 1995, Baillard et al., 1998, Girard et al., 1998). Thus most approaches have focused on reconstruction of specific building models: rectilinear shapes (McGlone and Shufelt, 1994, Roux and McKeown, 1994, Noronha and Nevatia, 1997, Collins et al., 1998), flat roofs (Berthod et al., 1995), or parametric models (Haala and Hahn, 1995, Weidner and Förstner, 1995). Recently, more generic reconstruction approaches involving multiple high-resolution images have been proposed (Bignone et al., 1996, Moons et al., 1998).

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. This paper presents our approach to solving this problem, which makes two contributions: First, an algorithm for matching individual line segments over multiple views; Second, an algorithm for computing planar facets of the scene starting from the matched lines. The key idea, which is common to both algorithms, is that both geometric and photometric constraints should contribute from all images.

In the line matching algorithm the match is computed using multi-view geometry and photometric similarity measures on the line neighbourhood in each image. Special attention is paid to using multiple images to overcome the deficiencies of the line segment feature extraction. In particular fragmented lines are joined and missing edges recovered. The algorithm is described in section 3.

In the plane generating algorithm 3D planar facets of the scene are computed by using both 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 method is in the use of inter-image homographies (plane projective transformations) to robustly estimate the planar facets. It 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 algorithm is described in section 4.

Our work is part of the European IMPACT project and we begin with an overview of this project.

2 THE IMPACT PROJECT

The IMage Processing for Automatic Cartographic Tools (IMPACT) project is a five site collaboration funded under the European Community Esprit Long Term Research Programme. The partners are the Universities of Bonn (Germany), Leuven (Belgium), Oxford (England), Ecole Nationale Supérieure des Télécommunications (ENST, France), and Eurosense (Belgium).

The objective of the IMPACT project is the automatic generation of 3D scene descriptions of urban and suburban areas from aerial imagery. Although this is a much studied problem in computer vision and photogrammetry, there are several aspects that differentiate the project from related work: First, the input data consists of high resolution colour images ($8.5 \times 8.5\ cm^2$ on the ground, RGB). Second, at least 3 images taken from different positions are available of the scene. The project goes further than the traditional 2-view stereo approach and exploits the redundancy in such image sets to the full. Third, the scenes are of higher complexity than those that have traditionally been studied. In European urban and suburban areas buildings are often packed together and of variable types and irregular shapes. Figure 1 shows an example of six typical overlapping views which are used as input data within the project.

The strategy explored within the project is to generate as quickly as possible 3-dimensional information and perform all grouping and modelling operations in the 3-dimensional world. There is an abundance of geometric primitives in the individual images, and great care has been taken that only *reliable* 3-dimensional information is generated. To achieve this, information is verified over multiple images with typically 4 to 6 images being used. Although there is a wide diversity in the shapes of the individual buildings, almost all of them can initially be described as polyhedral structures. Hence, *coplanarity* of points and lines is the dominant criterion present in the grouping and modelling algorithms.

The five partners have developed various parts of the integrated system. Often their areas overlap, but the following summarizes the main emphasis at each site.

Bonn focused on generating a symbolic 3D description of the scene from a number of symbolic 2D image descriptions. The 2D description consists of a Feature Adjacency Graph (FAG), for example a 2D corner and its associated line segments and regions. The idea is to transfer the neighbourhood relations that are available from the images into object space and to use them to produce consistent 3D object descriptions. This involves matching the 2D descriptions over multiple images. Details are given in (Brunn and Weidner, 1998, Fischer et al., 1998).

ENST developed methods for the identification of buildings in the images by refining and analysing dense disparity maps between image pairs. These methods are based on two major ingredients: an in-depth statistical study of the distribution of 3D points on homogeneous areas and a careful analysis of the geometrical properties of local planar surfaces. Photometric information and general knowledge about the scene geometry are also exploited to detect and to reconstruct the buildings. Details are given in (Fradkin et al., 1999a, Fradkin et al., 1999b).

Eurosense provided the source images used by all the partners. They are 11500 × 11500 pixel^2 aerial images, with a ground length of 8.5 cm per pixel. The calibration and the orientation of the camera were also supplied for each view. Six overlapping subparts are shown in figure 1.

Leuven focused on grouping 3D line segments into polygonal faces of the roof structure. This is achieved by first grouping the extracted 3D line segments into sets of coplanar line segments. For each such set initial polygon hypotheses are formed using an algorithm based on the convex hull of the line end points. These hypotheses are then verified both in 3D and by back-projection in the images. The emphasis during this modeling stage is on extracting the correct topology of the roof structure rather than on the metric accuracy of the reconstruction. This allows a more efficient roof modeling involving few criteria. Errors in the polygon's shape are corrected by exploiting symmetry relations and a general preference for rectangular structures in buildings. Metric accuracy is obtained in an additional step by back-projecting the recovered (wireframe) model of the roof structure into the images and minimizing the total reprojection error. Details are given in (Moons et al., 1998).

Oxford developed two areas. The first is a method for matching individual line segments between images. The method is described in section 3 of this paper and (Schmid and Zisserman, 1997). The second is a method for automatically reconstructing 3D piecewise planar models from multiple images of a scene. The method is described in section 4 and in (Baillard and Zisserman, 1999).

These approaches can cooperate in many ways. For example, the 3D lines produced by the Oxford line matching algorithm are used as input data by all partners. The strategy of ENST, based on dense disparity maps, is complementary to the feature-based approaches of the other partners. And the methods developed by Leuven and Oxford for generating polygonal faces are also complementary as will be explained in section 5.

3 LINE MATCHING OVER MULTIPLE VIEWS

Line matching is a difficult problem for several reasons. The first is due to the deficiencies in extracting lines and their connectivity: although the orientation of a line segment can be recovered accurately, the end points are not reliable, and furthermore the topological connections between line segments are often lost during segmentation. The second reason is that there is no strong disambiguating geometric constraint available over 2 views: In the case of points (corners), correspondences must satisfy the epipolar constraint. For infinite lines there is no geometric constraint, whilst for lines of finite length there is only a weak overlap constraint arising from applying the epipolar constraint to end points.

Existing approaches to line matching in the literature are of two types: those that match individual line segments; and those that match groups of line segments. Individual line segments are generally matched on their geometric attributes — orientation, length, extent of overlap (Medioni and Nevatia, 1985, Ayache, 1990, Zhang, 1994). Some such as (Crowley and Stelmazyk, 1990, Deriche and Faugeras, 1990, Huttenlocher et al., 1993) use a nearest line strategy which is better suited to image tracking where the images and extracted segments are similar.

Matching *groups* of line segments has the advantage that more geometric information is available for disambiguation. A number of methods have been developed around the idea of graph-matching (Ayache and Faugeras, 1987, Horaud and Skordas, 1989, Gros, 1995, Venkateswar and Chellappa, 1995). The graph captures relationships such as left of, right of, cycles, collinear with etc, as well as topological connectedness. Although such methods can cope with more significant camera motion, they often have a high complexity and again they are sensitive to error in the segmentation process.

The matching algorithm presented here is based on the one described in (Schmid and Zisserman, 1997), originally developed for three images. It is extended here to use any number of input images. The idea is to use the photometric neighbourhood of the line for disambiguation. Epipolar geometry is used to provide a point to point correspondence on putatively matched line segments over two images. The similarity of the line's neighbourhoods is then assessed by cross-correlation at the corresponding points. If the viewpoints are very different (for example a large baseline relative to depth or a substantially different attitude between

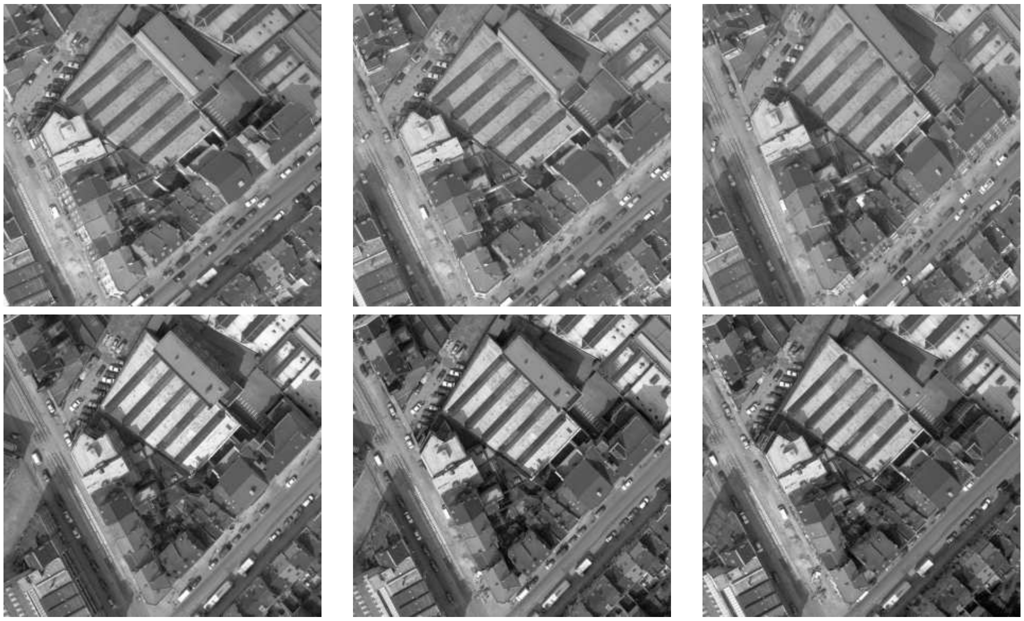


Figure 1: Six overlapping aerial views. The images are about 1200×1200 pixels, one pixel corresponding to a ground length of 8.5cm. The images are acquired at a height of 1300m in two triplets: the camera approximately translated by about 300m between successive views in images 1-3, and between images 4-6. The first set acquired on the outward flight, and the second set on the approximately parallel return flight.

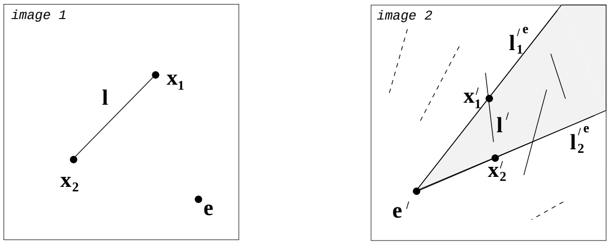


Figure 2: The epipolar geometry reduces the search space for line segments corresponding to $\mathbf{l}$. The segments have to lie in the region of the epipolar beam (Zhang, 1994) formed by $\mathbf{l}_1'^e$ and $\mathbf{l}_2'^e$, which are the epipolar lines corresponding to $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively.

views) then a correction is required to the cross-correlation neighbourhood. The computation of this correction (which is based on a planar homography) is described in (Schmid and Zisserman, 1997).

3.1 Two-view matching

This section describes the geometric and photometric constraints used in matching line segments between two views.

Reducing the search space. When matching lines over two views there is a weak overlap constraint for line segments of finite length arising from epipolar geometry. The two end-points of a segment generate two epipolar lines in the other image. These two lines define a region, called the epipolar "beam", which necessarily intersects or contains the corresponding segment (Zhang, 1994). Figure 2 illustrates this idea. Therefore a *search region* in the second image is determined for each segment in the first image. This reduces the complexity of the search for corresponding segments.

Pairwise matching score. For each pair of lines which satisfies the geometric constraint above, the photometric information available from the surfaces in the neighbourhood of the lines is used to compute a photometric matching score. The basic idea is to treat each segment as a list of points to which neighbourhood correlation is applied as a measure of similarity. Only the point to point correspondence is required, and this is provided by the epipolar geometry.

Corresponding image points, represented as homogeneous 3-vectors $\mathbf{x}$ and $\mathbf{x}'$, satisfy the epipolar constraint:

$$\mathbf{x'}^\top \mathtt{F}\mathbf{x} = 0.$$

$\mathtt{F}$ is the fundamental matrix, which is a 3×3 matrix of rank 2. The epipolar line corresponding to $\mathbf{x}$ is $\mathbf{l'}^e = \mathtt{F}\mathbf{x}$, and the epipolar line corresponding to $\mathbf{x}'$ is $\mathbf{l}^e = \mathtt{F}^\top\mathbf{x}'$. Note, lines are also represented by homogeneous 3-vectors, and $'$ in all cases indicates the second image.

Suppose two image lines, $\mathbf{l}$ and $\mathbf{l'}$ correspond (i.e. have the same pre-image in 3-space) then the epipolar geometry generates a point-wise correspondence between the lines. A point $\mathbf{x}$ on $\mathbf{l}$ corresponds to the point $\mathbf{x}'$ which is the intersection of $\mathbf{l}'$ and the epipolar line $\mathbf{l'}^e$ of $\mathbf{x}$: The point $\mathbf{x}' = \mathbf{l'} \times \mathbf{l'}^e = \mathbf{l'} \times (\mathtt{F}\mathbf{x})$ (see figure 3). This construction is valid provided $\mathbf{l}$ is not an epipolar line.

The matching score for a pair of line segments $\mathbf{l}$ and $\mathbf{l'}$ is computed as the average of the individual correlation scores for the points (pixels) of the line. The only points included in this average are those that are common to both line segments, i.e. the correlation is not extended past the ends of the measured segments. The pair hypothesis is valid if the matching score is above a certain threshold.

3.2 Three-view matching

With three views there is a strong geometric constraint available for line matching. The trifocal tensor (Spetsakis

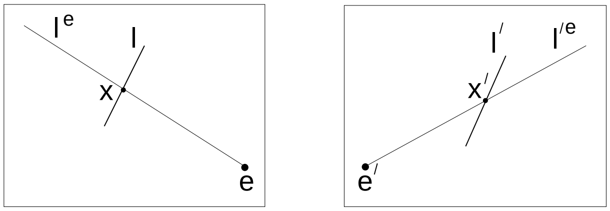


Figure 3: The point-wise correspondence between the lines $\mathbf{l}$ and $\mathbf{l}'$ is determined using epipolar geometry : $\mathbf{x}' = \mathbf{l}' \times \mathbf{l}'^e$ where $\mathbf{l}'^e$ is the epipolar line of $\mathbf{x}$.

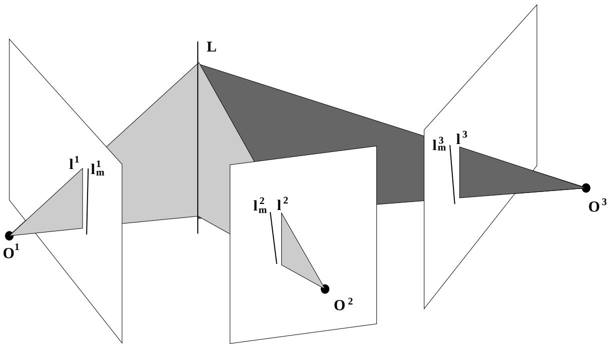


Figure 4: A line $\mathbf{L}$ is reconstructed in 3D by minimizing reprojection error over 3 views. In this case the error used is the distance between the end points of the measured lines, $\mathbf{l}^j_m$, and the image, $\mathbf{l}^j$, of the line $\mathbf{L}$ in each view.

and Aloimonos, 1990, Shashua, 1994, Hartley, 1995) enables lines matched in two views to be transferred to a third, and this process can be used to verify two view matches. This constraint provides a method for verifying putative pairwise line matches.

The matching algorithm over 3 views proceeds as follows: first a set of triple hypotheses is determined from putative pairwise matches, then consistent matches are selected.

Search for triple hypotheses. For every putative pair with a high correlation score (determined according to section 3.1), the position of the line in the third image is predicted using the trifocal tensor.

There are then two verification tests for the pair hypotheses:

- geometric verification: a line segment is detected at the predicted position in the third image ;
- photometric verification: there is a sufficiently strong pairwise correlation score between the second and the third image at the predicted line position.

There are two possibilities: If the geometric test is passed, the photometric test must also be passed, but the score threshold is more lenient than that used in the two view matching (about 50% of its value). Otherwise, if the geometric test is not passed, the pair hypothesis can still be verified by the photometric test, but the score threshold is more conservative than that used in the two view matching (about 50% more than its value). A match can therefore be verified whether or not there is a line segment at the predicted position.

A search is also computed proceeding from pairwise matches in views two and three and verifying in view one. Only consecutive views are used here for the two-view matching (i.e. matching is not performed between views 1 and 3) because of their greater photometric similarity.

Final matching set. The triple hypothesis set produced at the previous stage needs not to be consistent because one line can be associated with several lines in another image arising from different triplets. A consistent match set is therefore determined by a “winner takes all” scheme : whenever a line belongs to several triples, only the triple with the best score is retained.

Implementation details. The 2D line segments are extracted by a local implementation of the Canny edge detector at sub-pixel accuracy. Edgels are then linked into chains, jumping up to a one pixel gap. Tangent discontinuities in the chain are located using a worm, and line segments are fitted between the discontinuities using orthogonal regression. A very tight threshold is used for the line fitting so that curves are not piecewise linear approximated. Only line segments above a minimum length (here 15 pixels) are considered for matching. For computational efficiency, in each image the lines are stored in a raster format, i.e. a pixel indexes a line if there is one at that location.

The similarity score for a pair of line segments is computed as the average of the individual edgel correlation values. However, to be robust to occlusion, an individual correlation value is only included if above a threshold. Here a correlation window of 15×15 pixels is used and the threshold value is 0.6. If there are fewer than a minimum number of matched edgels for a putative segment match, then that match is eliminated. Here the minimum is 15 edgels.

If there are n lines in each image then the computational complexity of the 3-view matching algorithm would be $\mathcal{O}(n^3)$ if all lines are considered in each view. However, since triple hypotheses are formed from pairwise matches (and only verified in the third view) the complexity is $\mathcal{O}(n^2)$.

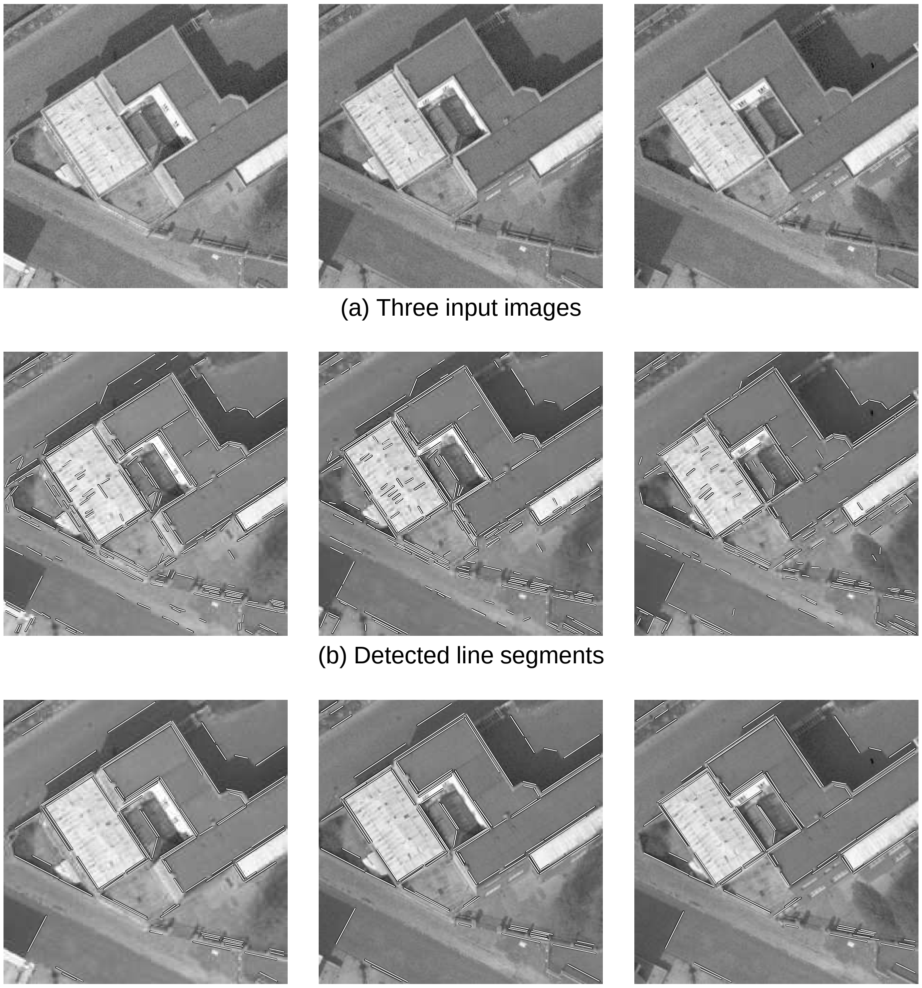
The lines are reconstructed in 3D by minimizing reprojection error over the 3 views using the Levenberg-Marquardt algorithm (see figure 4).

Results for 3-view line matching. For clarity the matching algorithm is illustrated on the 600×600 pixel^2 image set of figure 5. Figure 5b shows the detected 2D lines and figure 5c shows the 2D lines which have been matched. Figure 6a shows the corresponding 3D lines. About 40% of the detected lines are matched, and of these matches over 98% are correct. This is a very good performance: about 35% of the lines detected in one image are not detected in the other image, so although only 40% of the *detected* lines are matched, 62% of the lines which could be matched are indeed matched by the algorithm.

The matching results over 3 views for the data set of figure 1 are shown in figures 7a and 8a. About 20% of the detected lines have been matched.

3.3 Improving the quality of the matched lines

The line matches are refined in two ways. The first one – merging – overcomes a limitation of the winner takes all strategy. The second one – growing – overcomes the deficiency in line segmentation.

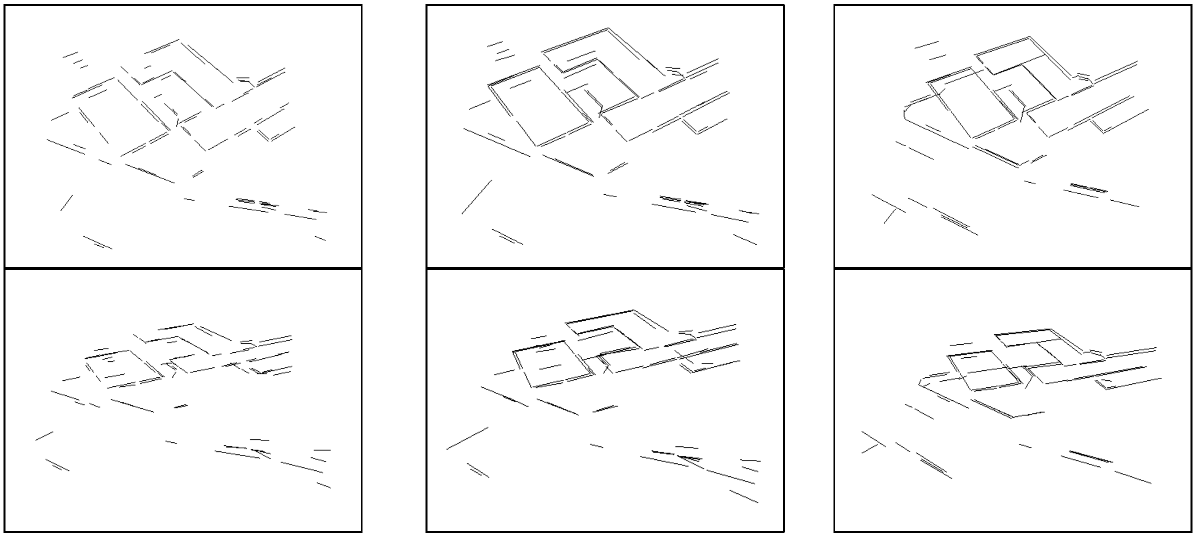


(a) Three input images

(b) Detected line segments

(c) Matched line segments

Figure 5: Three view line matching. The input images are 600×600 pixels. There are 248/236/212 detected line segments and 88 lines are matched, with only one erroneous match.

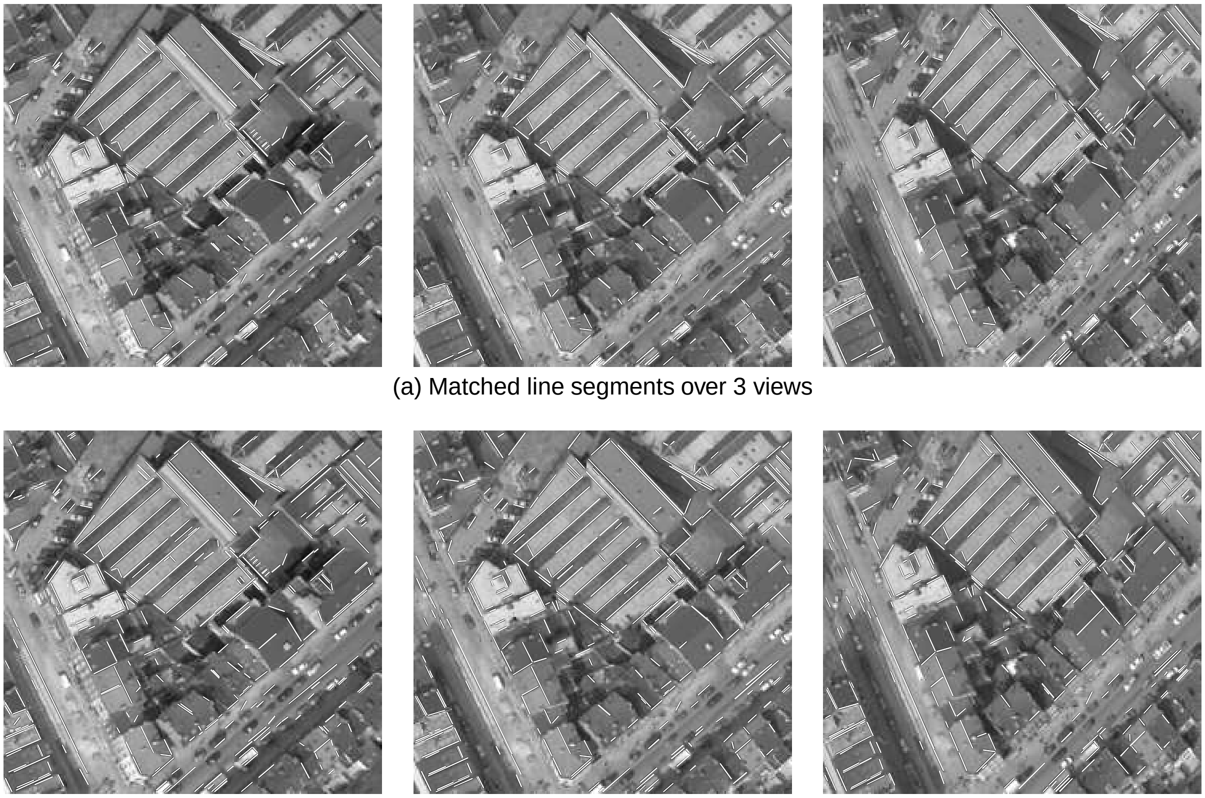


(a) 3D lines computed over 3 views before the merging and growing steps

(b) 3D lines computed over 3 views after the merging and growing steps

(c) 3D lines computed over 6 views

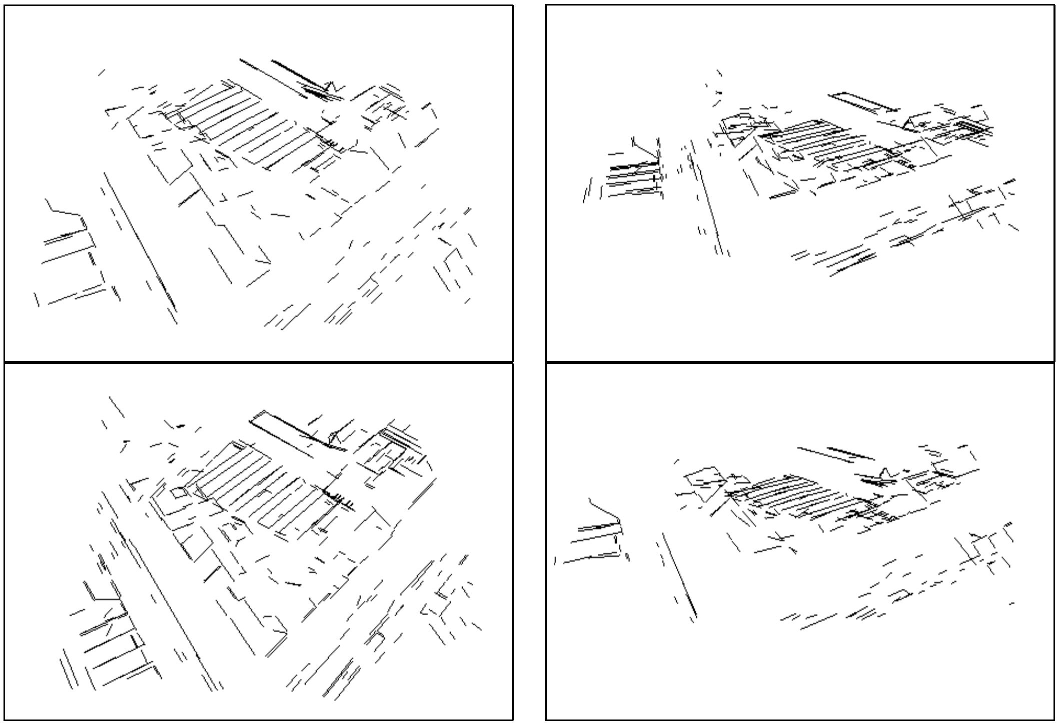
Figure 6: (a) and (b): Views of the 3D lines computed from the line matches of figure 5. The improvements of the merging and growing steps are clear: longer segments, shorter gaps, and better parallelism. The total length of the reconstructed 3D lines over 3 views is 473 m. (c): Views of the 3D lines computed over 6 views. There are 86 matched lines over 6 views, produced by merging together 88 and 80 triplets (a match is accepted if verified over 4 views at least). The total length of the 3D lines is 546 m. The quality has significantly improved compared to (b).



(a) Matched line segments over 3 views

(b) Matched line segments over 6 views

Figure 7: (a) Three view line matching for the first three images of figure 1. There are 1344/1307/1245 detected line segments and 229 matched lines (shown here). (b) Six view line matching for the images of figure 1. There are 332 matched lines over 6 views, produced by merging together 229 and 353 triplets. The figure shows the matched lines reprojected onto the 3 first images. Note, line segments missed in the triplet matches of figure (a) are recovered by matches in the second triplet.



(a) 3D lines computed over 3 views

(b) 3D lines computed over 6 views

Figure 8: (a) Two views of the 3D lines computed from the line matches of figure 7a, by matching over 3 views. (b) Two views of the 3D lines computed from the matched lines of figure 7b (by matching over 6 views).

Merging. During the line extraction, lines are often erroneously broken into several segments. Under "winner takes all" only one of these segments may be matched over three views. Suppose, for example, that in view one both lines $L1a$ and $L1b$ match line $L2$ in view two and line $L3$ in view 3. Even if lines $L1a$ and $L1b$ are collinear, only one of the two triplets $(L1a, L2, L3)$ and $(L1b, L2, L3)$ will be selected. If no further action is taken the effect of line fragmentation in detection is that line $L1b$ will be unmatched.

If $L1a$ and $L1b$ are collinear and the two triplets $(L1a, L2, L3)$ and $(L1b, L2, L3)$ are among the triplets before applying the winner takes all strategy, then it is likely that $L1a$ and $L1b$ arose from a single (but fragmented) line. The lines are merged if it is possible to fill in the gap between $L1a$ and $L1b$, that is if there is correlation support for the gap over 3 views. In this manner a line match is generated which is the *union* of the lines detected in each of the views. Furthermore 3D lines estimated from this correspondence are improved because longer line segments are available in each view.

Growing. Another deficiency of line detection, other than fragmentation, is that a fitted line may be shorter than the true line. However, the missing edgels are in general not the same in the different images. Thus if a short line is matched to a longer line (determined by epipolar lines), the end points may be extended provided the "grown" edgels satisfy the photometric correlation tests over the three views. This another significant benefit of using multiple views.

The significant benefits of this improvement are demonstrated when comparing figures 6a and 6b, which show the 3D lines computed before and after the line clean-up. About 15% additional line matches are generated by merging, and more than 70% of all matched lines are grown.

3.4 Extension to N-view matching

Using more views reduces matching errors, and increases the accuracy of the estimated 3D line. A natural method to extend the matching from 3 to N views would be to project the 3-view estimated line into subsequent images and verify its correspondence with a detected line in the new view using a combination of the geometric and photometric constraints described in section 3.2. Here a different method of extension is implemented which is better tuned to the acquisition circumstances of the image set.

As described in the caption of figure 1, the images were acquired in two sets. There are smaller photometric differences between images of the triple 1-3, and between images of the triple 4-6 than there are between any individual images. For this reason, lines are first matched within the triples, and then the triples are merged together.

In the case of 6 input images defining 2 triples, the matching is first performed independently for each. The correspondence between the triples is then established by reprojection in the images: Suppose $(\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3)$ is a line triple defined over views 1-3 and $\mathbf{L}_{123}$ the corresponding estimated 3D line segment. Similarly suppose $(\mathbf{l}_4, \mathbf{l}_5, \mathbf{l}_6)$ is a line triple defined over views 4-6 and $\mathbf{L}_{456}$ the corresponding 3D line segment. A correspondence is established between the two triples if the projection of $\mathbf{L}_{123}$ onto the images 4-6 matches at least two of the lines $\mathbf{l}_4, \mathbf{l}_5$, $\mathbf{l}_6$, and reciprocally if the projection of $\mathbf{L}_{456}$ onto the triple 1-3 matches at least two of the lines $\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3$. If a line triple has not matched any other triple then matches are sought from the set of unmatched lines in the images of the other triple. Whenever such a line is found, it is added to the matching set. A match has to be supported by a minimal number of views (here 4) to be valid. The method can easily be extended to $N > 6$ views by dividing the image set into triples. The triples may or may not overlap depending on the acquisition characteristics.

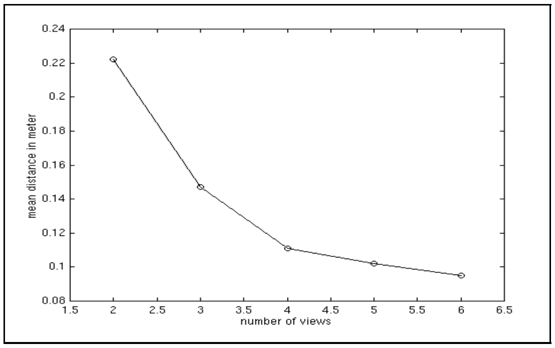


Figure 9: Mean height difference between endpoints of horizontal lines (from the images of figure 1) versus the number of views: the accuracy of the reconstruction increases with the number of views.

Results for N-view matching. Figures 6c, 7b and 8b show the matching results over 6 views for both example scenes, where a match has been accepted if verified over 4 views at least. The reconstructed 3D lines are of higher quality than those computed with only 3 views, and there are no mismatches at all. The total length of the reconstructed lines is about 15% higher than with 3 views. The accuracy of the reconstruction increases with the number of views, as shown in figure 9.

4 PRODUCING PIECEWISE PLANAR MODELS

This section describes the method for generating a piecewise planar model. The method proceeds from the computed 3D lines and consequently must cope with the shortcomings of that process and the earlier line detection: missing lines, fragmented lines, and the occasional mis-match.

The problems caused by missing features in piecewise planar reconstruction are illustrated by the detail in figure 13a, taken from the data set of figure 5. The correct roof model in this case is a four plane "hip" roof (Weidner and Förstner, 1995). 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 (Bignone et al., 1996, Moons et al., 1998), 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. The method described here requires only a single 3D line in order to generate a plane hypothesis. As will be seen it does not have the above failings of the two line methods.

The method consists of three main stages, which will be illustrated on the building shown in figure 13a:

1. *Computing reliable half-planes* defined by one 3D line and similarity scores computed over all the views (sec-

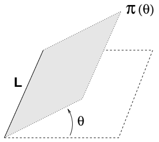


Figure 10: The one-parameter family of half-planes containing the 3D line $\mathbf{L}$. The family induces a one-parameter family of homographies between any pair of images.

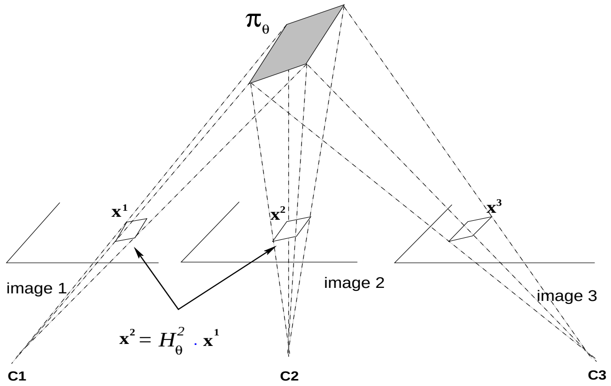


Figure 11: Geometric correspondence between views. θ, the homography $\mathtt{H}^i(\theta)$ determines the geometric map between a point in the first image and its corresponding point in image i.

tion 4.1). This is the most important and novel stage of the algorithm.

2. *Line grouping and completion* based on the computed half-planes (section 4.2). 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 4.3).

4.1 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 10). 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 hypothesizing 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 11. 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.

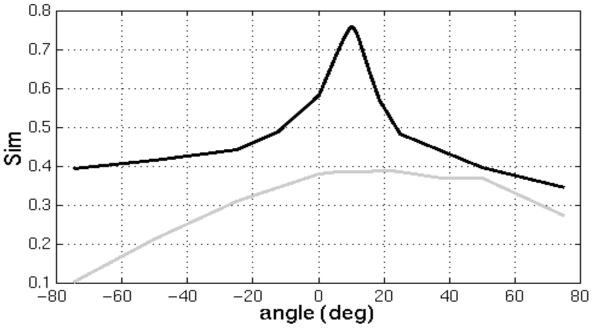


Figure 12: Example of similarity score functions $Sim(\theta)$. The black curve corresponds to a valid plane, whereas the grey one is rejected. The following validity criteria are used: maximum value of the function $Sim(\theta^{max}) \geq 0.4$, absolute value of the estimated second derivative around the maximum $|Sim''(\theta^{max})| \geq 4.0$, and global amplitude $Sim(\theta^{max}) - Sim(\theta^{min}) \geq 0.2$.

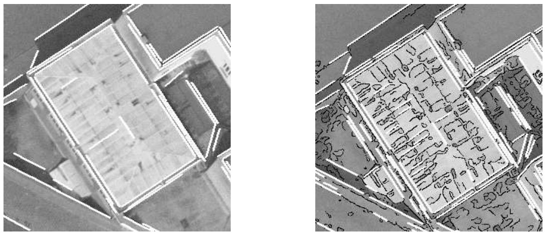

(a) (b)

Figure 13: (a) Detail of figure 5a 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.

In more detail, the plane $\boldsymbol{\pi}(\theta)$ attached to a 3D line $\mathbf{L}$ is assessed by a similarity score computed over the image set 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 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 $\boldsymbol{\pi}$ in the world frame, then (Luong and Vieville, 1996)

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

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:

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

and ranges between $(0, 1)$. Figure 12 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

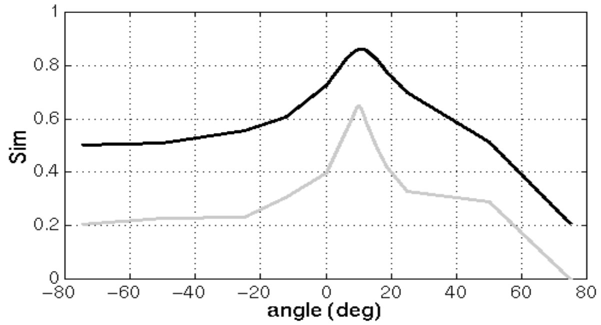


Figure 14: Effect of the baseline on 2-view similarity scores. The black curve corresponds to a short baseline between views (e.g. views 1 & 2), the grey curve to a wide one (e.g. 1 & 4). The curves apply to the same half-plane. A short baseline leads to high maxima (low distorsion between the images), 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.

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 computed by applying an edge detector to the first image, with a very low threshold on gradient (an example is given in figure 13). The edges are then linked and regularly sampled to determine the POI. The choice of the first view is arbitrary and can be automated by selecting the most textured image.

The correlation term $Cor(\mathbf{x}, \mathbf{x}^i)$ is the centred normalized 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 even more selective.

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 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 14) in the data set.

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 12. 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. In the following the parameter values for line matching have been relaxed: a minimal number of only 10 matched edgels (rather than 15) is required to verify a line match, in order to produce more half-plane hypotheses. This is possible because erroneous hypotheses can be culled all along the process.

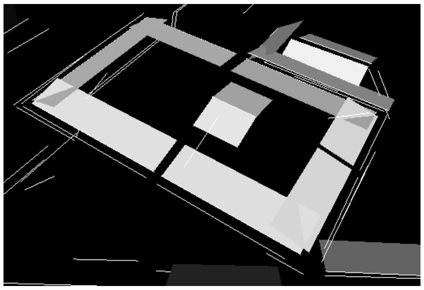


Figure 15: Detected half-planes over the interval $[-75^o; +75^o]$.

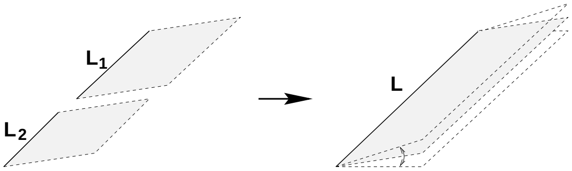


Figure 16: Collinear grouping. The optimal plane angle is recomputed for the merged line, again using $Sim(\theta)$ as described in section 2. This is more accurate than, for instance, averaging angles.

Figure 15 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.

4.2 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.

Collinear grouping (figure 16). Two collinear lines which have attached coplanar half-planes are merged together. The result of the collinear grouping of half-planes of figure 15 is shown in figure 18a.

Coplanar line and half-plane grouping (figure 17). Any line which is neighbouring and coplanar with the current plane is associated with it (see the example of figure 18b).

Creating new lines by plane intersections (figure 19). 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 20).

4.3 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 lines* (boundaries). The initial support line of a plane is a natural border line. Additional border lines are created as shown in figure 21.

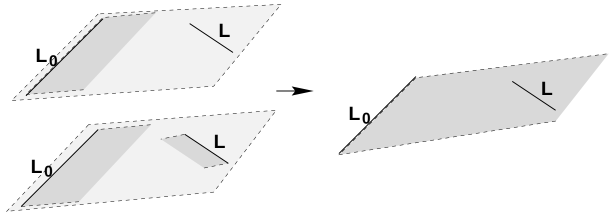


Figure 17: Coplanar line and half-plane grouping. In the top case, $\mathbf{L}$ belongs to the half-plane $\boldsymbol{\pi}(\mathbf{L}_0)$, and a new plane is computed by orthogonal regression to a regular point sampling of $\mathbf{L}_0$ and $\mathbf{L}$. In the bottom case, $\mathbf{L}$ has an attached but consistent half-plane, therefore the two plane hypotheses are merged into a new unique plane, also computed by orthogonal regression.

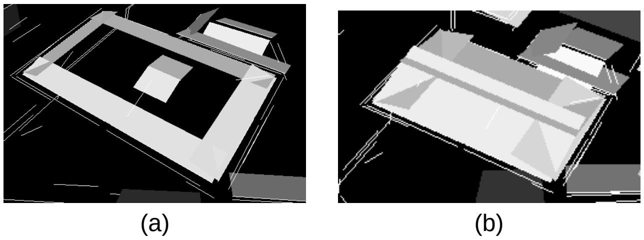

(a)
(b)

Figure 18: 3D 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.

A closed delineation can then be computed by using heuristic grouping rules (Weidner and Förstner, 1995, Noronha and Nevatia, 1997, Moons et al., 1998) to associate border lines. For instance the end points of the border lines are updated when lines intersect or have close end points (see figure 22). Then the convex hull of the border line end points is computed. 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.

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. Figure 23 shows both the 2D delineation and a 3D view of the roof produced for the building of figure 13a.

Finally, occlusion prediction is used to signal and resolve conflicts between inconsistent plane hypotheses. A conflict occurs between two facet hypotheses when their projections onto an image substantially overlap, i.e. when one of them is occluded by the other. Where conflicts between plane hypotheses arise, there are two possibilities, depending on the coplanarity of the conflicting planes.

If the the conflicting planes are coplanar, then the planes are merged together. The position of the new single plane is computed using lines belonging to both merged planes, and its delineation is obtained by merging the previous delineations together.

On the contrary, if the the conflicting planes are not coplanar, then the conflict is resolved using a confidence score $S(\mathcal{P})$, which denotes the quality of the face $\mathcal{P}$:

$$S(\mathcal{P}) = \mathcal{A}(\mathcal{P}) \times \tau_{real}(\mathcal{P}),$$

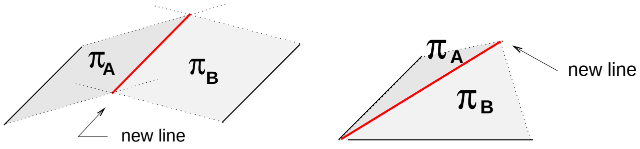


Figure 19: Creation of new lines when two planes intersect.

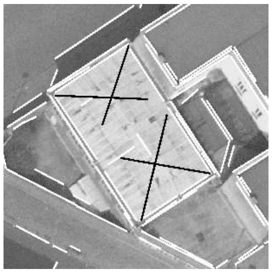


Figure 20: New lines (black) created by plane intersection.

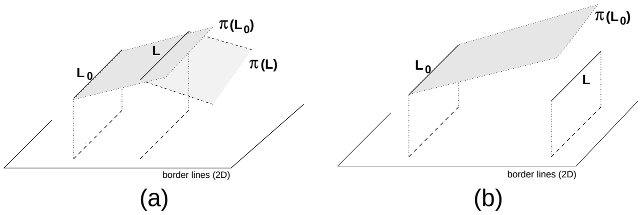
where $\mathcal{A}(\mathcal{P})$ is the area of the face, and $\tau_{real}(\mathcal{P})$ the ratio of the length of input 3D lines to hypothesised lines in the delineation. The plane hypothesis with the lowest confidence score is suppressed.

Whenever a plane is merged or removed, the neighbouring planes are updated. In grouping operations, thresholds on distances are avoided by using a topological neighbourhood between projected lines, defined by a Delaunay triangulation constrained to fit the line segments. This also enables quick access to neighbours.

4.4 Results of model building

Figure 24 shows the 3D reconstruction of the full scene of figure 5a. Figure 25 shows the result on the much larger and more complicated images of figure 1. Note that intricate and unusual roofs (for example the factory in the upper part of the image) have been completely recovered. This also demonstrates how little photometric texture is required by the method, since roofs with virtually homogeneous intensity are retrieved. Only two roofs are missed in the entire scene.

Performance. The quality of the reconstruction is governed by the completeness and correctness of the input line set. The method is robust to a proportion of missing and erroneous lines because mechanisms are included to generate new lines, by plane intersection, and to cull erroneous lines with their associated half-planes, in the final verification stages. However, the performance is improved if too many, rather than too few, lines are supplied. This is because a line is the only mechanism for instantiating a



(a)
(b)

Figure 21: Border line computation for plane delineation: (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.

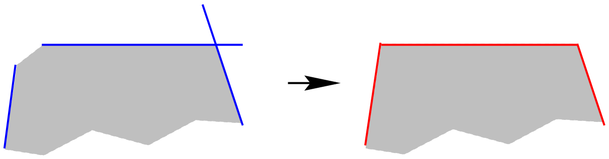


Figure 22: **Updating end points of the border lines**: any end point outside the region of interest is moved into it; two close endpoints are replaced by the intersection of the two lines

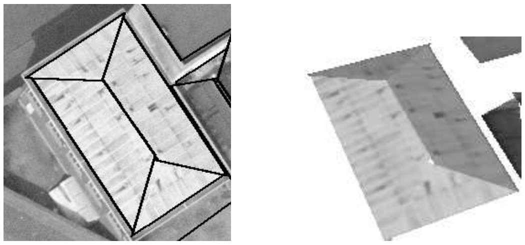


Figure 23: Example of a reconstructed roof. (a) Delineation of the verified roofs projected onto the first image; (b) 3D view with texture mapping.

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 13), 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.

5 FUTURE WORK AND EXTENSIONS

The results demonstrate that it is possible to automatically construct models of urban scenes from multiple images using quite minimal information. The models are of very reasonable quality.

How might the quality be improved further? The most significant improvement would be to provide additional features for matching. For example, a bar detector would gain extra lines in the example building of figure 13. A second improvement would be the incorporation of other, albeit more restrictive, 3D grouping mechanisms. The approach of this paper is not an alternative to roof generation systems based on the grouping of at least two neighbouring coplanar lines (Bignone et al., 1996, Moons et al., 1998), but is complementary to such systems. An efficient and robust approach would use both, grouping and removing from consideration coplanar lines and using the described single 3D line method for the remainder. This requires an architecture for a cooperating strategy to be developed.

It would also be preferable if the similarity score function had a firmer statistical foundation. Measurements on the function correctly facilitate plane rejection and angle computation, but the measures are 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.

Acknowledgements. Financial support for this work was provided by the EC Esprit Project IMPACT and the UK EPSRC IUE Implementation Project GR/L05969.

REFERENCES

Ayache, N., 1990. Stereovision and Sensor Fusion. MIT-Press.

Ayache, N. and Faugeras, O., 1987. Building a consistent 3D representation of a mobile robot environment by combining multiple stereo views. In: Proc. IJCAI, pp. 808–810.

Baillard, C. and Zisserman, A., 1999. Automatic reconstruction of piecewise planar models from multiple views. In: Proc. CVPR.

Baillard, C., Dissard, O. and Maître, H., 1998. Segmentation of urban scenes from aerial stereo imagery. In: Proc. ICPR, pp. 1405–1407.

Berthod, M., Gabet, L., Giraudon, G. and Lotti, J. L., 1995. High-resolution stereo for the detection of buildings. In: A.Grün, O.Kübler and P.Agouris (eds), Automatic Extraction of Man-Made Objects from Aerial and Space Images, Birkhäuser, pp. 135–144.

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

Brunn, A. and Weidner, U., 1998. Hierarchical bayesian nets for building extraction using dense digital surface models. Int. Journal of Photogrammetry and Remote Sensing 53(5), pp. 296–307.

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

Crowley, J. and Stelmazyk, P., 1990. Measurement and integration of 3d structures by tracking edge lines. In: Proc. ECCV, pp. 269–280.

Deriche, R. and Faugeras, O., 1990. Tracking line segments. In: Proc. ECCV, pp. 259–267.

Fischer, A., Kolbe, T. H., Lang, F., Cremers, A., Förstner, W., Plümer, L. and Steinhage, V., 1998. Extracting buildings from aerial images using hierarchical aggregation in 2D and 3D. CVIU 72(2), pp. 185–203.

Fradkin, M., Roux, M. and Maître, H., 1999a. Building detection from multiple views. In: ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery.

Fradkin, M., Roux, M., Maître, H. and Leloglu, U., 1999b. Surface reconstruction from multiple aerial images in dense urban areas. In: Proc. CVPR. to appear.

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

Gros, P., 1995. Matching and clustering: Two steps towards object modelling in computer vision. Intl. J. of Robotics Research 14(6), pp. 633–642.

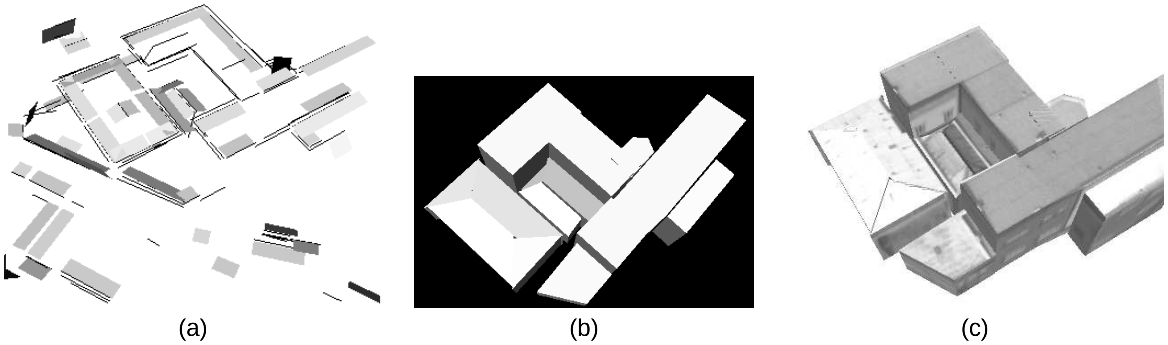


Figure 24: Model reconstruction results on the full example scene of figure 5. (a) 49 detected half-planes from 137 3D lines. (b) 3D model of the scene (12 roof planes). The vertical walls are produced by extruding the roof's borders to the ground plane. (c) 3D model of the scene with texture mapping

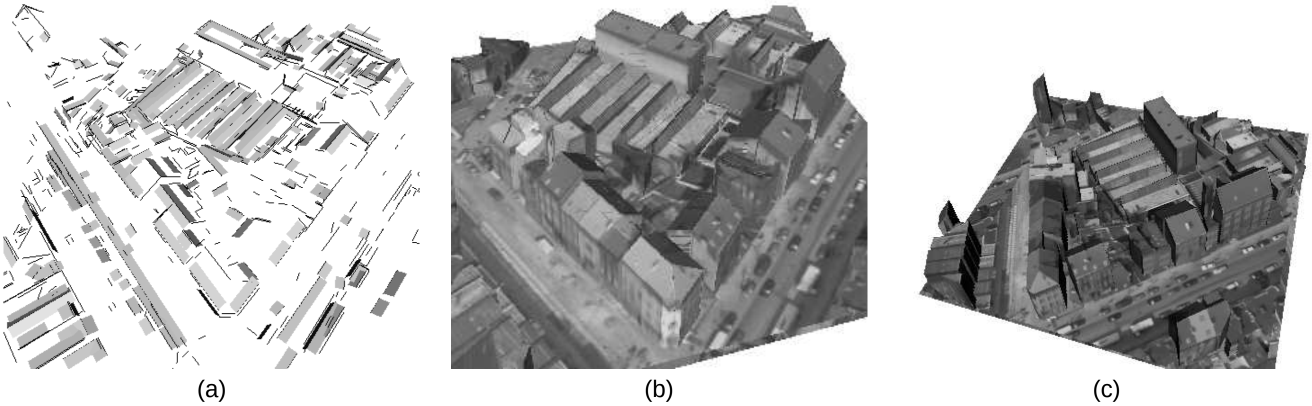


Figure 25: Model reconstruction results on the large (1200×1200 pixels) images of figure 1. (a) The 452 reconstructed 3D lines and the 267 detected half-planes. (b) and (c) Two views of the 3D model of the scene, with texture mapping (180 roof planes).

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

Hartley, R. I., 1995. A linear method for reconstruction from lines and points. In: Proc. ICCV, pp. 882–887.

Horaud, R. and Skordas, T., 1989. Stereo correspondence through feature grouping and maximal cliques. IEEE T-PAMI 11(11), pp. 1168–1180.

Huttenlocher, D. P., Klanderman, G. A. and Rucklidge, W. J., 1993. Comparing images using the Hausdorff distance. IEEE T-PAMI.

Luong, Q. T. and Vieville, T., 1996. Canonical representations for the geometries of multiple projective views. CVIU 64(2), pp. 193–229.

McGlone, J. and Shufelt, J., 1994. Projective and object space geometry for monocular building extraction. In: Proc. CVPR, pp. 54–61.

Medioni, G. and Nevatia, R., 1985. Segment-based stereo matching. Computer Vision, Graphics and Image Processing 31, pp. 2–18.

Moons, T., Frère, D., Vandekerckhove, J. and Van Gool, L., 1998. Automatic modelling and 3d reconstruction of urban house roofs from high resolution aerial imagery. In: Proc. ECCV, pp. 410–425.

Noronha, S. and Nevatia, R., 1997. Detection and description of buildings from multiple images. In: Proc. CVPR, pp. 588–594.

Roux, M. and McKeown, D. M., 1994. Feature matching for building extraction from multiple views. In: Proc. CVPR.

Schmid, C. and Zisserman, A., 1997. Automatic line matching across views. In: Proc. CVPR, pp. 666–671.

Shashua, A., 1994. Trilinearity in visual recognition by alignment. In: Proc. ECCV, Vol. 1, pp. 479–484.

Spetsakis, M. E. and Aloimonos, J., 1990. Structure from motion using line correspondences. IJCV 4(3), pp. 171–183.

Venkateswar, V. and Chellappa, R., 1995. Hierarchical stereo and motion correspondence using feature groupings. IJCV pp. 245–269.

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

Zhang, Z., 1994. Token tracking in a cluttered scene. Image and Vision Computing 12(2), pp. 110–120.