COMPUTER VISION AND IMAGE UNDERSTANDING
Vol. 72, No. 2, November, pp. 122–142, 1998
ARTICLE NO. IV980722

Building Detection and Reconstruction from Mid- and High-Resolution Aerial Imagery

N. Paparoditis

Equipe Traitement des Images et du Signal (ETIS), Université de Cergy–Pontoise, 6 avenue du Ponceau, F 95014 Cergy–Pontoise Cedex, France; and Ecole Supérieure des Géomètres Topographes (ESGT), LOEMI, 2–4 avenue Pasteur, F 94165 Saint-Mandé Cedex, France*
E-mail: paparoditis@icare.ign.fr

and

M. Cord, M. Jordan, and J.-P. Cocquerez

Equipe Traitement des Images et du Signal (ETIS), Université de Cergy–Pontoise, 6 avenue du Ponceau, F 95014 Cergy–Pontoise Cedex, France
E-mail: {cord, jordan}@ensea.fr; cocquerez@u-cergy.fr

Received April 15, 1997; accepted December 20, 1997

In this paper, we discuss methods for building detection and reconstruction from aerial imagery. These methods are intended for the analysis of urban and suburban areas and have been applied to images of different resolutions (between 1 m and 10 cm per pixel). Various algorithms for image matching have been investigated, including hierarchical processing and new correlation schemes that have interesting properties for building recognition and building feature grouping. Cooperative combination of 2-D (monocular) and 3-D (stereoscopic) information allows the complete representation of the observed scene and particularly the detection of man-made raised structures such as buildings. A performance evaluation on simulation-based images has been considered in comparison with the corresponding ground truth reference. Our work illustrates that mid-resolution methods cannot be directly applied to high-resolution images. Classical algorithms must be adapted and new techniques have been defined to carry out dense urban area reconstruction. © 1998 Academic Press

INTRODUCTION

Over the past few years, automatic and semiautomatic extraction of cartographic features from aerial and spatial images has become a subject of intensive research. Man-made features, such as road networks and buildings, have been given special attention [16, 17].

This research takes place in a context where there is an increasing need for updated maps, particularly in urban areas, requiring the use of digital and, as far as possible, automatic cartographic processes. Applications particularly concern environmental modeling and urbanism.

Stereophotogrammetry is the basis of most methods presented in the literature, although some results can be obtained by single image analysis (shadow analysis [24, 32]) or fusion of range data and aerial images [19]. Digital stereophotogrammetry consists of six successive steps [3]: (1) image acquisition and digitization, (2) geometric modeling, (3) image primitive detection, (4) image primitive matching, (5) 3-D reconstruction, and (6) final 3-D surface reconstruction and modeling.

This paper deals with a set of methods that we have developed in order to extract buildings from images of urban and suburban areas. These methods are applied to stereo pairs of digital or digitized aerial and satellite images whose resolution is between 1 m and 10 cm per pixel. The exterior orientation of the images is supposed to be known and the methods concern steps (3) to (6) of stereophotogrammetry. Primary versions of our system were developed for the reconstruction of suburban scenes essentially composed of flat-roof or planar-roof buildings. So, we used a polyhedral scene model for building reconstructions. In later developments, we intend to use as little *a priori* knowledge as possible concerning the nature and shape of buildings: roof shapes are planar and buildings are supposed to have linear edges, without any assumption about corner angles.

In Section 1, we present related research that is similar to our approach. Section 2 presents two different systems we developed in order to extract 3-D information and especially man-made structures in suburban areas from *mid-resolution* (about 1 m per pixel) aerial images. The first system aims at detecting buildings from a complete disparity map obtained by various stereo correlation techniques, whereas the second system is based on a prior monocular interpretation of images (which provides building detection) followed by a 3-D computation restricted to particular points in the images i.e., building corners). Section 3 presents some recent work we have developed for *high-resolution* (about 10 cm per pixel) images representing urban areas.

* ETIS is an associated research unit of french CNRS (URA 2235).

1077-3142/98 $25.00
Copyright © 1998 by Academic Press
All rights of reproduction in any form reserved.

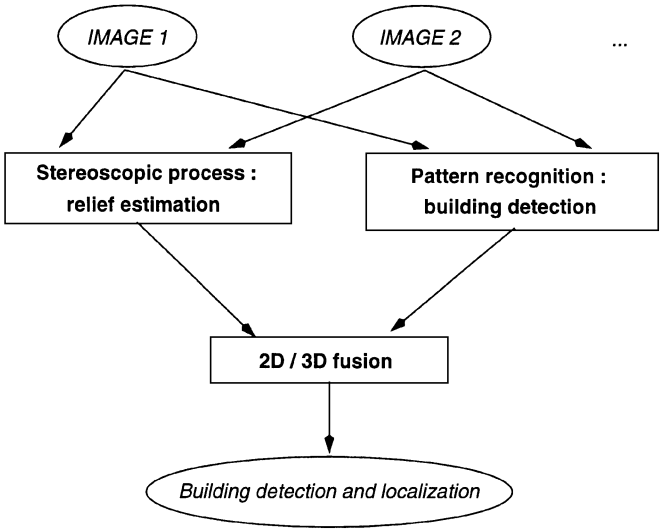


FIG. 1. General building detection scheme.

1. RELATED WORK

Building detection and recognition from aerial images is such a complex problem that it cannot be solved in a single step [43]. On the contrary, it must be solved by using a combination of different techniques involving

- stereoscopic processing, which is itself a complex research field,
- monocular image processing, particularly involving pattern recognition, image interpretation, shadow analysis, and structural image analysis.

A very general scheme of these processes is shown in Fig. 1.

The purpose of this section is not to provide a complete overview of the domain, but to mention the work which has influenced our approach.

Kim and Muller [29, 30] describe a system combining stereopsis and 2-D information extraction: stereopsis generates a 2.5-D surface by pyramidal region growing matching; building hypotheses are developed from line detection and lines relations search; then the fusion step consists of the interpolation of height data from stereoscopic process in the boundaries of building hypotheses from monocular process.

Lotti and Giraudon [34] have developed an approach similar to Kim *et al*. The originality of their method lies in the matching process [33], see Fig. 2.

Matching is accomplished by a correlation method using an adaptive size window adjusted on the contours of both images. Different methods are used to extract contours from each view. The windows are set up as follows: for each matching candidate point, four windows corresponding to four directions are built. For each direction, the maximal recovery surface between the "west" and "east" views, constrained by contours, is selected. Then the largest window from all directions is named the "minimal window." It is selected if and only if it verifies a boolean relation taking into account three predicates based on texture, noise, and a thresholding of correlation. Finally, the correlation is computed on the basis of these "minimal windows." Then the "west" and "east" disparity are densified with a growing process working in four-connectivity and with an interpolation process [28]. Moreover, a pyramidal approach has been implemented to reduce computational coast. Results on epipolar aerial images having a ground pixel size about 40 cm are claimed to have a mean error of about 0.4 m and rms error of about 1.7 m.

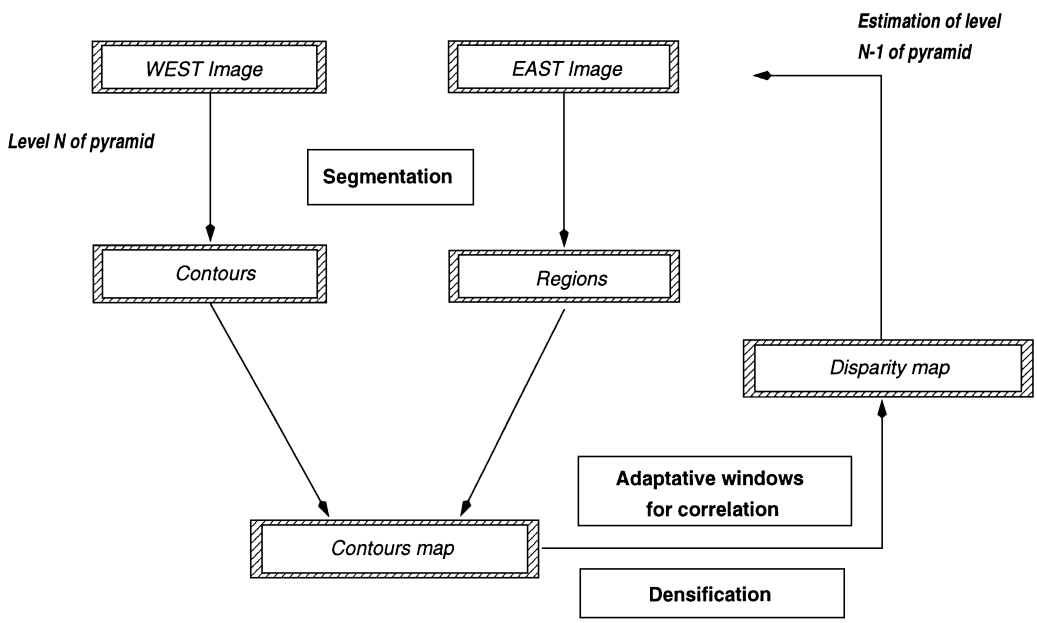


FIG. 2. Lotti and Giraudon's 3-D reconstruction system.

The *ASCENDER system* developed at the University of Massachussets aims at modeling rectilinear and flat-roof isolated or connected buildings. The data set consists of mid-resolution aerial images with known exterior orientation. Best results are obtained with one nadir and 4 oblique (E, W, N, and S) views, but a combination of other images may be used [10, 25].

The first step in the process is the monocular detection of buildings by linear edge detection, corner extraction, and perceptual grouping. False detections or gaps can be corrected by multiimage analysis. Detected roof vertices are sought for in all images and a 3-D wire-frame representation of the scene is built up. Building roofs are extruded down to the ground that is itself modeled by a DTM [44], and the complete scene can be visualized by projection of the image radiometries on the 3-D representation. More than 80% of the buildings are correctly detected and the planimetric and altimetric precision is claimed to be less than 1 m.

Weidner and Förstner developed an approach for extracting 3-D building shapes from a digital elevation model (DEM), using constraints on building models [46, 47]. DEM are computed using image and feature pyramids and the final surface is then refined by means of local adaptive regularization techniques; this process gives a graph surface $z = z(x, y)$. Building detection is based on the fact that buildings are higher than the topographic surface, which is estimated using mathematical morphology on the DEM. The window size used in this process requires prior knowledge of the expected maximum size of buildings in the scene. Buildings are reconstructed depending on their complexity. Two kinds of parametric models are used for simple buildings, having either a flat or symmetric sloped roof. For complex or connected buildings, prismatic models are used.

The aim of the *AMOBE project* at ETH Zürich is the 3-D reconstruction of man-made objects from aerial images [21]. The data set consists of four vertical color aerial images, having a ground resolution of about 10 cm and with precisely known sensor orientation. Figure 3 presents the project's scheme.

DEM generation is performed by a standard correlation algorithm and provides a DEM with grid spacing of between 25 and 50 cm. 3-D blob detection is performed on the DEM and combined with color image information in order to extract building hypotheses. Feature extraction consists of an edge detection and an aggregation step, followed by the computation of photometric attributes on edges. 2-D segments are extracted from one image and corresponding features are sought for in the other images by maximizing an "edginess function" which is computed on all pixels along the epipolar line. Grouping of 3-D segments is performed on the four images and provides a set of 3-D planes, which are assembled in order to model building roofs.

Other systems use less stereoscopic information and combine monocular information such as shadow analysis or incremental multiimage processing.

Roux and MacKeown [41] presented a system for the construction of 3-D roof surfaces of buildings by integrating additional information from multiple view analysis. The basis of this system is the detection of 3-D corners using two views to extract polygonal surfaces from corner and edge cycles in a graph. This process involves the use of geometric constraints in the object space. As in the MOSAIC system [22], the analysis of new images increases the 3-D accuracy of surfaces and adds potential new buildings that were not perceptible in first views. The results show the efficiency of the system with various viewpoints (nadir and oblique) and its complementarity with other approaches [35].

Contrary to this iterative multiple image strategy, Wiman and Axelsson [48] used parallel search strategy. Image features are accumulated in a common frame in object space and analyzed in this frame. The two main horizontal directions in object space are extracted from six images. Then, a 3-D clustering of the line segments that contribute to these directions is carried out and the most reliable 3-D lines are merged to obtain 3-D planes. The ground resolution is about 12 cm. At the moment, results are limited to simple images.

Recent work by Chung and Nevatia deals mostly with shape issues and is applied to monocular images; the 3-D analysis deals with 3-D object descriptions using shadows to obtain 3-D information [6]. The final step is a hypotheses generation process mostly based on the work described in [37]. Given a set of basic shapes, pertinent structural relationships for the domain can be deduced by decomposing the basic shapes into the structural relationships. In this work, the case of object shapes (building

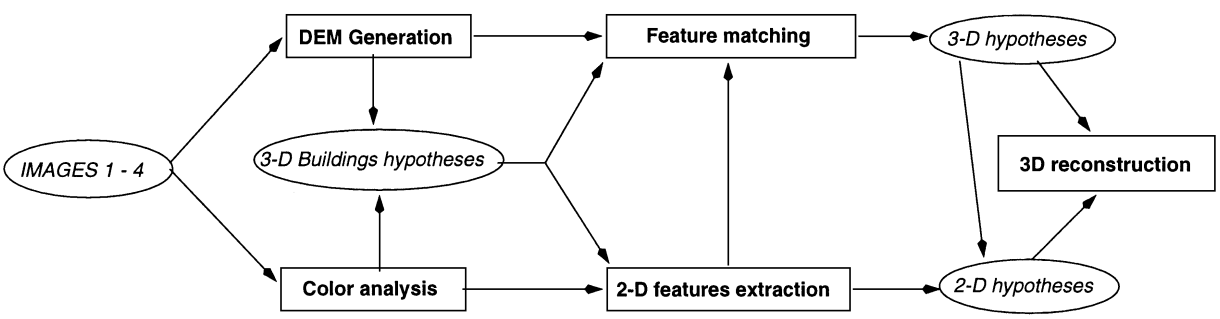


FIG. 3. ETH Zürich's AMOBE project.

roofs) composed of rectangles has been considered. The system uses a feature hierarchy that encodes the structural relationships specific to this set; lines, parallels, U-contours, and rectangles are identified as the pertinent structures into which these shapes can be decomposed. After the formation of all reasonable rectangles, a selection process is used to choose rectangles which have strong evidence of support, avoiding potential conflicts. In this system there are two kinds of rules: local selection rules and global selection rules. Local selection rules determine whether or not a rectangle is regular based on the local supporting evidence of the rectangle; the goal of the global selection rules is to select a minimum set of rectangles which could best describe the composition of the scene. Verification and generation of 3-D models use shadows for the validation and the estimation of height extracted from monocular views; both assume that the sun angle is given.

2. RECONSTRUCTION OF SUBURBAN AREAS FROM AERIAL IMAGERY

In this section, we present two methods for the reconstruction of suburban areas from mid-resolution aerial imagery. These methods were initially intended for the representation of low–density suburban areas essentially containing rectilinear and flat-roof large buildings, such as industrial or apartment house areas. The two approaches were simultaneously developed with our laboratory's further projects.

The first approach (Section 2.1) is based on a 3-D surface reconstruction of the scene performed by various stereo correlation techniques. Building extraction is then accomplished by filtering this 3-D surface.

The second approach (Section 2.2) is rather different and is based on a prior building recognition step by means of a monocular image interpretation. Aerial monocular image interpretation is another research field in our laboratory and an image understanding system composed of a connexionist classifier and a constraint propagation algorithm have been developed [9]. 3-D coordinates of each building are then computed by a matching step based on projective geometry.

For both approaches, the data set consists of a stereo pair of aerial or satellite images, whose resolution is between 80 cm and 1 m per pixel. Relative and absolute orientations are known, so that image disparities can be converted into cartographic 3-D coordinates.

2.1. Approach 1: Scene and Building Reconstruction Based on Global Image Matching

The first approach consists of the following steps:

1. computation of a 3-D surface model from the stereo pair by means of various correlation techniques;
2. extraction of buildings from this 3-D surface, using basic knowledge and assumptions on the building 2-D and 3-D shapes.

These techniques have been applied to various images, both real and simulated from aerial and satellite systems with a ground resolution of around 1 m.

2.1.1. Real and Simulated Data

2.1.1.1. Real data. Real data are extracted from an aerial image pair supplied by I.G.N. (Institut Géographique National) and covering the city of Cergy, France. The images have been digitized, so that the pixel ground resolution is about 85 cm.

We will illustrate initial steps of the method in Figure 7 by an example of processing. Figures 7a and 7b show, respectively, "west" and "east" images of the pair. In epipolar images such as these, the epipolar lines are confused with the scanlines, and the disparity (or parallax) between two corresponding points of the images (representing the same point of the scene) is reduced to a difference of pixel column indices.

2.1.1.2. Mapping technique evaluation based on simulated data. The automatization of 3-D mapping techniques comes up against a major difficulty, viz. the validation or the performance evaluation of the mapped products [23, 39]. This difficulty is mainly linked to the lack of reference data. When reference data are available and reliable, an external validation can be considered. This consists of measuring the discrepancy between the map obtained and that of the reference. This method, which is the most frequently used, comes up against two types of obstacles. First, reference data are very seldom adapted to this type of validation, either because they are insufficient in number (thus leading to statistics which are devoid of meaning) or because they are themselves marred by errors. Then, a definition of the distance between the map and the reference is not trivial. Indeed, it must be pertinent from a statistical point of view and reflect the type of information that the map is supposed to provide (e.g., position, altitude, shape, and orientation). Furthermore, the evaluation of the quality of the mapping with respect to a reference has meaning only if the reference is of a much higher quality than the mapped product.

A reference obtained by manual stereo plottings at image scale will not support a good evaluation of the reconstruction because even if they are of subpixel quality in the image grid, they may be marred by errors that can be too great for the desired subpixel accuracy of the measurement. A quantitative evaluation on nonrealistic synthetic images is none the more representative. These images are only good enough to give a behavioral idea of the tested algorithms but yet do not allow a performance assessment on complex realistic images like aerial or satellite images. In an attempt to overcome these problems, a simulation-based performance evaluation method is proposed. Realistic image simulation, based on landscape modeling and geometric and radiometric sensor modeling, offers a wide range of advantages for the validation of mapping techniques. First, it permits the multiplication of situations, i.e., landscapes, on the one hand, and optical systems on the other. One can thus try to test algorithms in multifarious and even extreme situations. This variety of situations

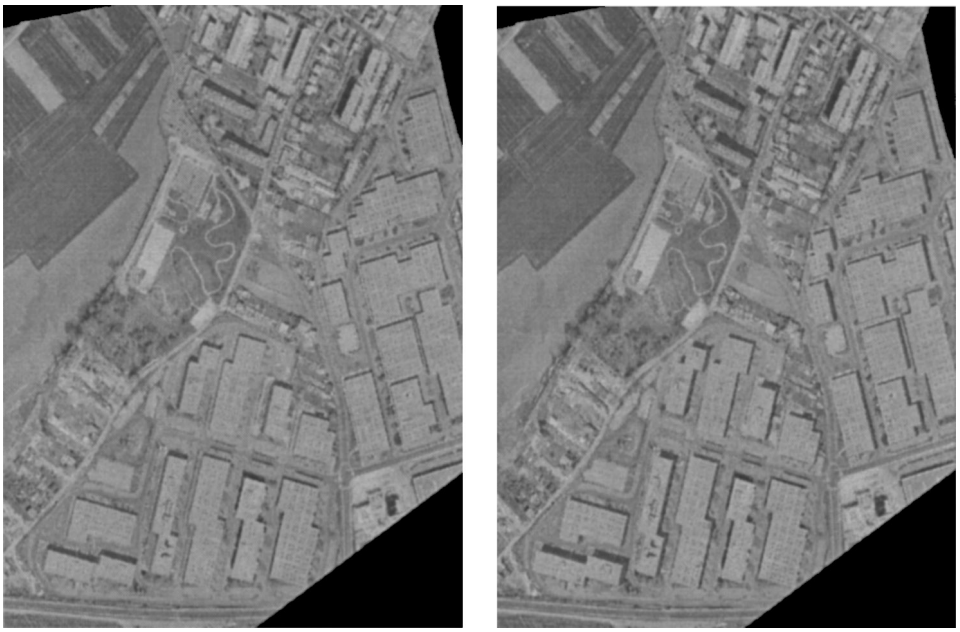


FIG. 4. Digital satellite simulated stereo pair of a suburban scene.

can be treated in an entirely parametric way. The impact of each parameter (whether linked to the system or to the ground) on the system performance can thus be studied separately by varying it while keeping all the others fixed. Finally, the simulation approach allows a quantitative parametric validation between the precise and dense "ground truth" reference data, given by the landscape model, and the mapped data.

Figure 4 shows a realistic simulated satellite side-looking stereopair (ground resolution = 1 m, MTF = 0.22, SNR = 200, -17° and $+17^\circ$ viewing angles). We computed it with a simulator developed by *Aérospatiale* adapted to cartographic targets such as buildings [45]. The simulator's landscape input is a "high-frequency" data base describing the urban superstructures in the form of a 3-D TRAPU model provided by the I.G.N. and a set of aerial photographs of the same zone. This simulated lansdcape (the town of Rungis in the Paris suburbs) is particularly interesting because of the large variety of building sizes and shapes. We can point out industrial buildings, appartment houses, and suburban houses. One can notice that some of the buildings are virtual (painted on the ground). Virtual buildings are helpful to test the efficiency of our building recognition scheme.

From the landscape model, one can build any kind of reference maps needed for mapping techniques evaluation. Figure 5a shows a reference DSM in image space. It has been computed by back-projection of the 3-D TRAPU model faces into the sensor's image plane by a z-buffer technique. The corresponding (X, Y, Z) for each image pixel is then calculated by an interpolation in the corresponding 3-D model face. A reference disparity map is computed in the same way. Figure 5b shows the building roof edges' reference. Both the reference maps shown clearly confirm that some buildings are virtual.

2.1.2. Image Matching and Surface Reconstruction

Image matching is the main problem of stereovision. The key issue is to obtain the highest number of matches, with the fewest incorrect correspondences.

Many basic image primitives can be used for stereo matching: pixels, edgels, linear segments, and regions. They have all been investigated, leading to various matching algorithms and properties. In order to reduce the combinatorial aspect of matching, constraints are often used; the search can be restricted along epipolar lines and the disparity interval can be estimated from the expected relief variation in the scene; unicity (*one primitive in the first image has at most one corresponding primitive in the other image*) and order (*primitives have the same order in both images*) constraints can be used.

We have developed two different techniques using contour features to constrain image matching and surface reconstruction.

2.1.2.1. Hierarchical matching approach and surface reconstruction using radiometric regions. In order to detect relief discontinuities and obtain a dense matching map, we have chosen a hierarchical approach applied to epipolar images:

1. edge point detection and matching, supposed to be reliable for relief transition detection;
2. cross correlation constrained by the previous edge matching and applied to the other points.

This approach is briefly described in the following paragraphs and summarized as shown in Fig. 6. Interested readers can refer to [27] for more details.

Edge detection and matching. The edge detector used in this step is the one proposed by Canny and Deriche [14]; contour points are those where the response to the filter is the highest. Successively applied with different values of window size α, the filter allows the characterization of contour points (local contrast) and enables the definition of an optimal value α_{opt} of α at each point (value for which the contour is delocalized). Edge detection is followed by two-level thresholding on gradient norms.

Both images of the pair are independently processed. Results are shown in Figs. 7c and 7d. The gray-level assigned to edge pixels in these images correspond to the value of α_{opt} for these pixels. Edge detection was first performed with $\alpha = 2.0$ and then with decreasing values down to $\alpha = 0.2$ by a 0.1 step.

The algorithm for edge point matching is as follows:

1. The knowledge of the absolute altitude interval in the scene allows the computation of the disparity interval Δ_d.
2. For each edge point P_w in the "west" image, search for a corresponding edge point in the "east" image: the search interval is the interval of length Δ_d centered in the point having the same coordinates as P_w. The corresponding point is that which, if it exists, is characterized by the closest value of α_{opt}.
3. When the entire image has been processed, matching consistency is tested on interscanline contours; this allows the elimination of isolated or aberrant matchings.

This algorithm relies on the assumption that relief discontinuities are perceptible by similar radiometric discontinuities in both images. The user must provide a threshold Δ_α to define the similarity between edges, i.e., the maximum difference between the values of α_{opt} for two corresponding edge points. The experiments that we carried out show that the tuning of this parameter is not a key point of the algorithm.

Figure 7e shows the results of edge point matchings. Δ_α was 0.4. The disparity interval was $[-10, +10]$. Black or white pixels indicate whether the disparity is positive or negative. Gray pixels are nonmatched pixels.

Cross correlation. By localizing relief discontinuities, edge matching allows identification of structures, especially man-made structures. Nevertheless, the number of matched points is too small to allow a correct surface estimation.

In order to increase the number of matched points, the second step of the matching process is a classical correlation step computed inside the regions defined by relief discontinuities.

The cross-correlation function F is computed on rectangular $(2u+1) \times (2v+1)$ windows

$$F(\xi) = \frac{1}{(2u+1)\cdot(2v+1)} \times \sum_{\iota=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} \frac{W(\iota,\eta)\cdot E(\iota,\eta+\xi) - m_W(\iota,\eta)\cdot m_E(\iota,\eta+\xi)}{\sigma_W(\iota,\eta)\cdot\sigma_E(\iota,\eta+\xi)}, \tag{1}$$

where $W(\iota,\eta)$ and $E(\iota,\eta+\xi)$ are the gray levels in both images, m_W and m_E are the mean gray levels in windows and σ_W and σ_E are the gray level standard deviations. Points close to contours, and such as the correlation windows would overlap the contour, are not considered.

For each point of the "west" image, the algorithm is defined as follows:

1. an *estimated corresponding point* in the "east" image is computed from the position of the left and right closest matched contour points;
2. a *search window*, centered at this point, is defined in the "east" image, and the correlation function is computed by varying ξ inside the window;
3. the matched points are those for which $F(\xi)$ presents a unique maximum, greater than the threshold which the user provides.

In fact, for simplicity reasons we always use square correlation windows. The tuning of the window size parameter must be carried out carefully, the window must be sufficiently large to enclose significant structures and sufficiently small to avoid relief smoothing. All the tests that we made on various images used 5×5 to 11×11 windows. Figure 7f shows the result with 7×7 windows. The same gray scale coding is used for disparity values in Fig. 7e.

The following step of the processing is the conversion of these disparity values in absolute elevation values by means of both relative and absolute orientation models.

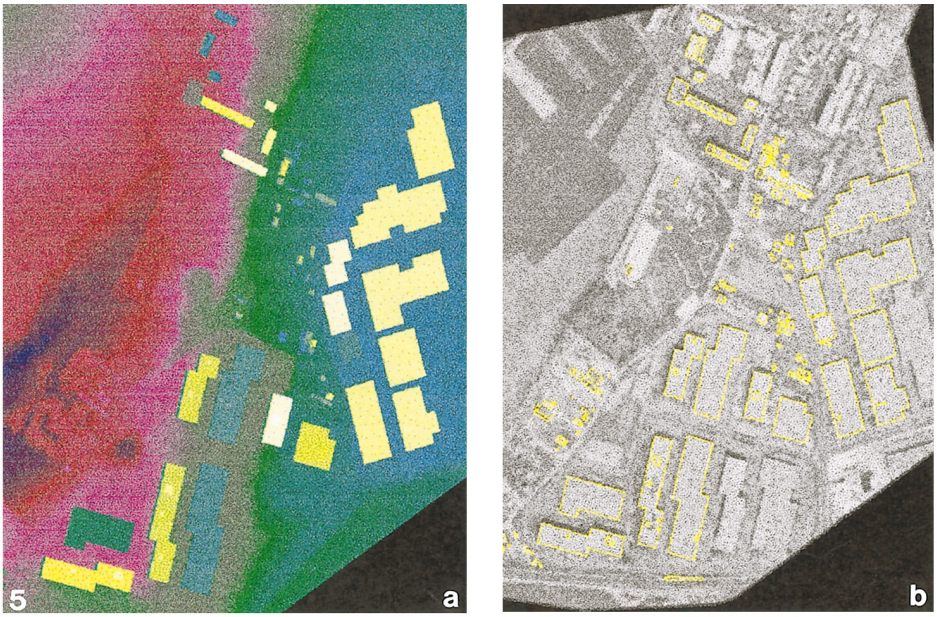
Surface reconstruction: Using a radiometric segmentation to compute mean 3-D planes. The elevation samples obtained from the image matching step are sparse and need to be densified. We first tried to obtain a complete model of the scene by combining a radiometric segmentation of one image with the sparse 3-D data obtained as described in the previous section. The image segmentation is a very classic one based on edge detection and contour closing.

The algorithm is then composed of the following steps:

1. the computation of a 3-D mean plane, either horizontal or oblique, for each region of the segmentation; each region is described by an equation $z = ax + by + c$;
2. the merging of adjacent regions that are represented by planes having close values of (a, b, c);
3. the splitting, along gradient maxima lines, of regions where the 3-D data are too sparse to allow plane modeling and reiteration of the plane modeling for these regions;
4. the insertion of vertical sides at the frontiers between altimetric regions.

A planar model of the scene is obtained, which is particularly suited to the representation of suburban areas.

Figure 7g shows an image segmentation of the image shown in Fig. 7a. Figure 7h shows the elevation map resulting from surface reconstruction. The brightest areas are those with the highest altitude.



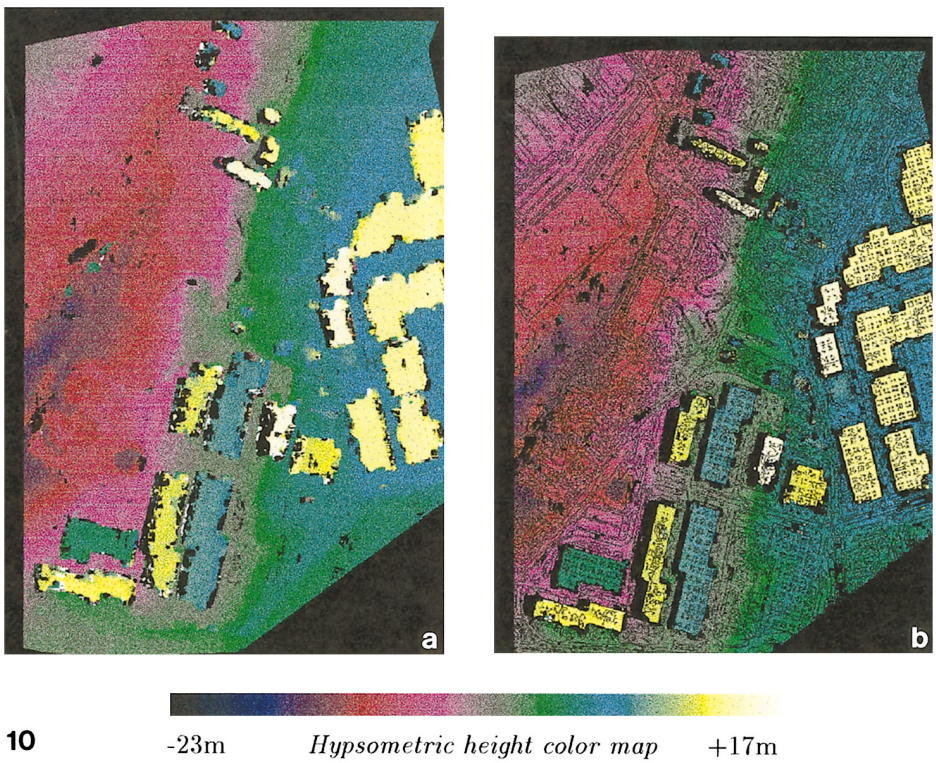


FIG. 5. Reference data: (a) reference DSM, (b) building edge reference.

FIG. 10. DSM results: (a) height map obtained by classical template matching and (b) height map obtained by edge preserving mask matching.

Some more results. Figure 8 shows the results obtained for a stereo pair (600 × 400) with a ground resolution of 85 cm per pixel.

The altitude map was computed from a 1115 closed-regions segmentation; about 40% of the pixels are matched.

Tests made on different scenes show that the altimetric accuracy of the method is about 1.5 m on man-made structures. Other examples are provided in [27].

The main drawback of this process is its dependence on low-level segmentation quality and on the structures radiometric

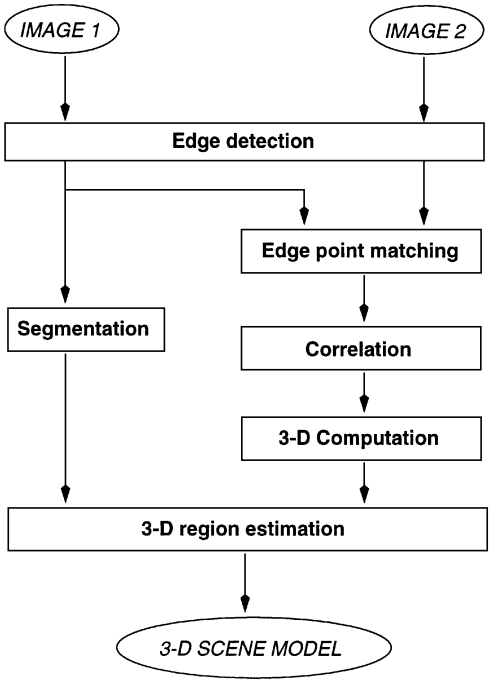


FIG. 6. Scheme for the hierarchical approach.

contrasts, and in the following paragraph we present another correlation method that we developed to avoid this difficulty.

2.1.2.2. Edge preserving template matching approach and surface reconstruction. One of the major problems of classical template matching techniques is the inherent smoothing of the mapped surface particularly across elevation discontinuities. Template matching is incapable of rendering well-located and clean building elevation discontinuities. Indeed, the disparity/elevation associated with a pixel is a "mean" (depending on radiometric distribution) of all pixel disparities/elevations in the template window. To reduce this smoothing effect the window sizes are usually reduced. In turn, the reduction of the texture sample leads to a rise in the correlation noise and in the number of mismatches.

This contradiction and the limitations of a fixed window size for all the images has encouraged the development of adaptive template window size techniques [28, 33]. The Okotumi and Kanade correction technique consists of an *a posteriori* correction of an initial set of disparity values determined by a classical cross-correlation technique for a given window size. The new window size is fixed by an analysis of the radiometric and disparity distribution of the window content.

The results are efficient only when the initial disparity set is good. The Lotti and Giraudon technique is based on an *a priori* use of 2-D image information to guide the choice of the window sizes. They are constrained by the contour points extracted from both images. The chosen window for each pixel is the biggest rectangular window not overlapping a contour. This technique gives the most efficient results. Nevertheless, results get sparser as the contour density rises and the disparity accuracy is affected by the fact that template windows are not centered on the mapped pixel.

Principle of new adaptive window matching. Edge-preserving mask correlation is a very simple template matching

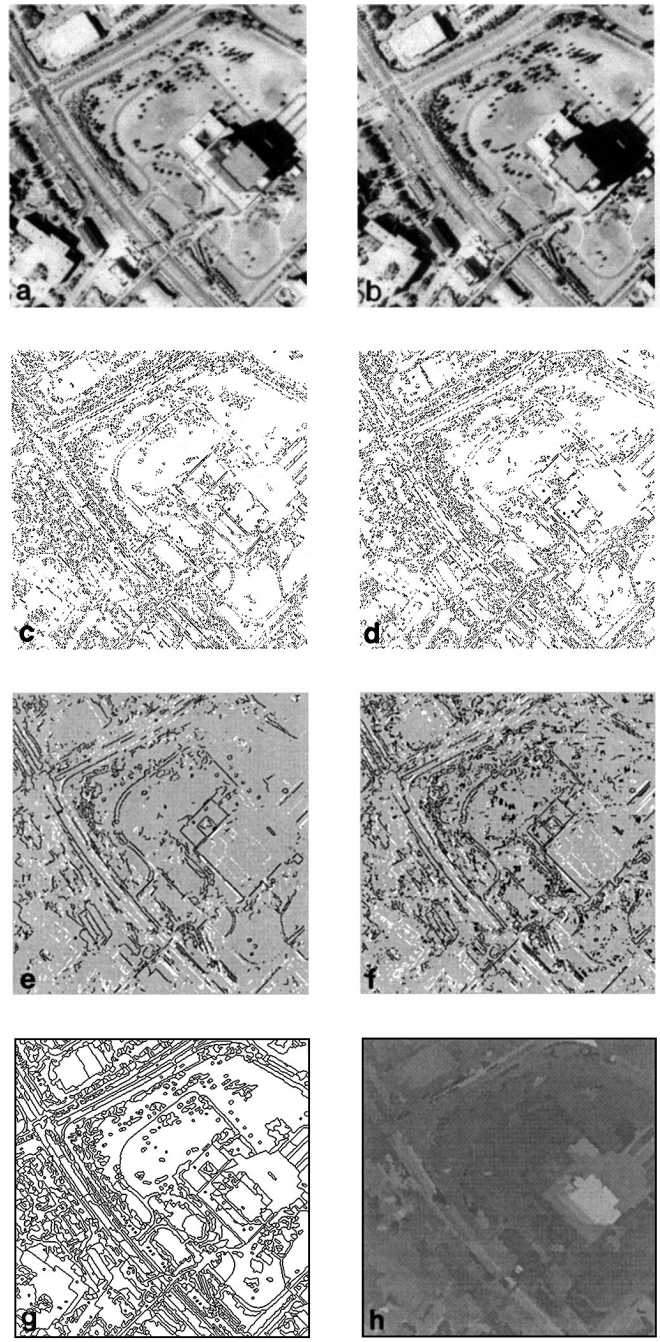


FIG. 7. Running example of the hierarchical approach. (a) "west" image, (b) "east" image, (c) edge "west" image, (d) edge "east" image, (e) edge point matching, (f) nonedge point matching, (g) radiometric segmentation, and (h) altitude map.

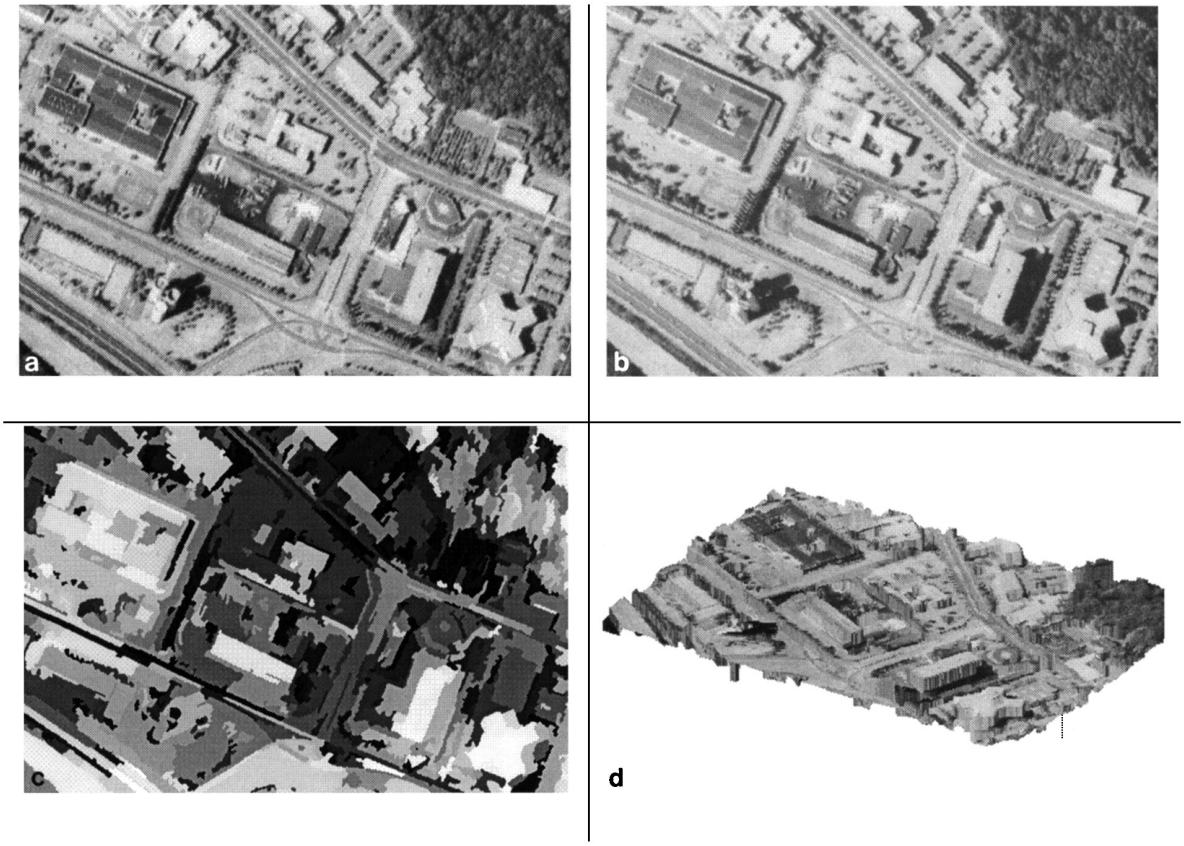


FIG. 8. Using a radiometric segmentation to compute mean 3-D planes: (a) "west" view, (b) "east" view, (c) altitudes map, and (d) a perspective view of the scene.

technique using adaptive windows [38]. It is based on the hypothesis that height transitions and slope breaks are, most of the time, characterized by radiometric transitions in the image space given by a contour map obtained by any good edge detector.

each pixel (i, j), we determine by a 4-neighbor region growing in window *win* the binary mask S giving all the pixels on the same side of an image contour as pixel (i, j). The new edge preserving mask correlation score can be easily expressed

$$F'(\xi) = \frac{\sum_{\iota=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} S(\iota, \eta) \cdot ((W(\iota, \eta) \cdot E(\iota, \eta + \xi) - m'_W(\iota, \eta) \cdot m'_E(\iota, \eta + \xi))/(\sigma'_W(\iota, \eta) \cdot (\sigma'_E(\iota, \eta + \xi)))}{\sum_{\iota=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} S(\iota, \eta)}, \qquad (2)$$

The idea of our technique is to restrict the calculation of the classical cross-correlation template matching score to some pixels of the template window. These pixels belong to the form or to the region described by all the pixels connected to the center pixel of the template that are on the same side of the contour. This prevents the smoothing disparity effects due to the use of pixels on the other side of 2-D contours, which could well be 3-D depth discontinuities. The classical template matching algorithm then becomes a new matching algorithm halfway between an area-based and a feature-based matching algorithm. (See Fig. 9.)

New correlation score. Let *win* be the template window extracted from the map contour corresponding to the radiometric template window of pixel (i, j) of one of the two images. For

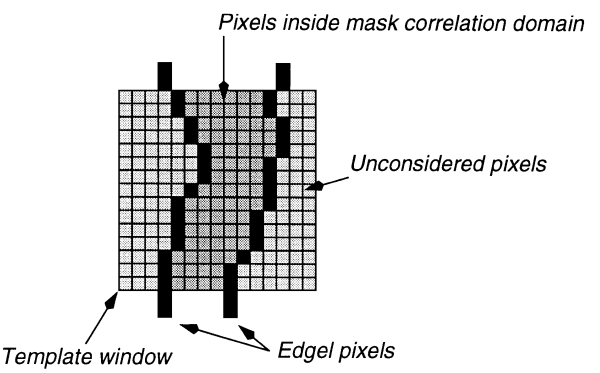


FIG. 9. Mask correlation principle.

where

$$S(\iota, \eta) = \begin{cases} 1 & \text{if } (\iota, \eta) \in R(i, j) \\ 0 & \text{otherwise,} \end{cases}$$

where $R(i, j)$ is the region formed by all the pixels connected to pixel (i, j) which are not contour pixels and $m'_E, m'_W, \sigma'_E, \sigma'_W$, respectively, are the new means and variances computed on all samples of the template belonging to $R(i, j)$.

Advantages and limitations of this algorithm. The disparity grids obtained are dense and accurate. Furthermore, the disparities are edge preserved and fill the grid up to the contours. This new template matching technique is quite an improvement in comparison with classical cross correlation. In comparison with adaptive window size techniques constrained by contours, the mask correlation technique gives better results on high-density contour maps. Nevertheless, adaptive form and adaptive size window matching are not in competition, they are complementary. Adaptive window size techniques can be very helpful for the matching of very homogeneous regions (low texture) often occurring in urban images.

Subpixel disparity estimation. A local interpolation of the correlation surface allows the improvement of the disparity precision to a subpixel measurement. The correlation peak is modeled by a polynomial approximation of the Shannon interpolator using the 4×4 neighbors around the discrete maximum. The subpixel peak position is estimated by a gradient conjugate descent method.

A subpixel two-way filtering technique. Pixels located in occluded areas, in periodic patterns, or in low textured areas are frequently mismatched. To avoid those mismatches, a two-way confirmation filtering technique [26] is applied to the entire disparity image. This filtering technique studies the disparity coherence between the disparity map computed with the "east" image as the slave image and the disparity map computed with the "west" image as the slave image. We have adapted this technique to subpixel disparity precisions which strengthen the filtering efficiency. The pixel (i, j) of the "east" image is discarded if the following expression is not valid

$$\inf_{n\in[0,1], m\in[0,1]} \| \overline{V_E}(i, j) - \overline{V_W}(\lfloor i + V_{Ei}(i, j) + n \rfloor, \lfloor j + V_{Ej}(i, j) + m \rfloor) \| \leq \tau, \tag{3}$$

where $\overline{V_E}$ (respectively, $\overline{V_W}$) is the disparity vector for a pixel of the "east" (respectively, "west") image, V_{Ei} (respectively, V_{Ej}) is the row (respectively, column) component of this vector. τ defines a good matching quality estimator. The mismatches usually follow a uniform distribution. A one-way matching is considered valid if its disparity error is smaller than twice the mean value of standard deviation matching errors estimated (by simulation-based evaluation, see Section 2.1.2.3) on several stereo-pair samples ($2 \times 0.3 = 0.6$ pixels). Thus, a two-way matching is considered valid if the expression (Eq. (3)) is valid with $\tau = 2 \times 0.6 = 1.2$ pixels.

DSM results. The elevation corresponding to the computed disparities can be computed with the steroscopic calibration knowledge. Figure 10a shows the DSM in image space computed by a classical cross-correlation technique and Fig. 10b shows the DSM in image space computed with our new adaptive window matching technique with $\tau = 1.2$ and with the very dense Canny–Deriche contour map ($\alpha = 3.0$) shown in Fig. 13a. Both have been computed with 15×15 window sizes. The DSM obtained by our new technique is dense, precise, and reliable. We think that merging adaptive size window and adaptive window form matching techniques could provide even more reliable disparity maps or DSM. However, the DSM quality is dependent on the effectiveness of the edge detection method used. For elevation discontinuities characterized by low-contrast transitions, missing in the contour map, the resulting mask will be the whole template. Therefore, the depth discontinuity will be smoothed.

2.1.2.3. Image matching and DSM quality estimation. The total errors estimated in comparison with a reference, due to mapping, is the sum of errors due to the sensor's geometrical calibration and errors due to matching. An additional advantage of the simulation is that one can simulate an infinitely accurate sensor to uncorrelate the errors of calibration and of matching. The remaining matching errors integrate the matching intrinsic errors and the errors due to generalization, which are linked to a scale difference between the images used for mapping and the images used to obtain the reference. Therefore, to overcome these generalization errors, the shapes of the described details of the superstructures are properly chosen in order to elude the problem of scale dependence of shape generalization as much as possible.

The performance evaluation results of the disparity map and the corresponding DSM, shown in Tables 1–3, have been obtained by a point control method. It consists of evaluating the quality of the mapped data from corresponding samples known simultaneouly in the mapped data and in the reference data. The quality assessment of a corresponding pair of points is estimated by their distance (the Euclidean distance for instance). For a set of corresponding points, the global quality assessment is given by the rms error if the reference is accurate.

TABLE 1
Matching Precision (Pixels)

	Matching precision (pixels)		
Themes	Row RMS	Column RMS	Total RMS
All	0.17	0.21	0.27
Ground	0.16	0.22	0.28
Roofs	0.15	0.18	0.25

TABLE 2
DSM Theme Precision

Themes	DSM precision (meters)	
	Planimetric RMS	Altimetric RMS
All	0.19	0.36
Ground	0.19	0.37
Roofs	0.18	0.30

The altimetric errors, just as matching errors, follow a Gaussian distribution (Fig. 11a). A close study of the disparity errors shows that estimated disparities are affected by a systematic error depending on the disparity value. As shown in Fig. 11b, this bias, due to the subpixel disparity interpolator, is a function of the fractional part of disparity.

2.1.2.4. Surface reconstruction. The elevation samples are sparse. A complete surface reconstruction must go through a densification of elevation data by interpolation techniques. Because of all the occluded areas, the elevation interpolation process should never be carried out in image space. The DSM data should be first resampled in object space. Nevertheless, a blind interpolation process of the missing data, using local existing grid neighbors, can be dangerous as we do not know where elevation discontinuities really occur and how the elevation surface texture behaves. So its seems necessary on the one hand to restrict the used neighbors to the properties belonging to the same 3-D structure and on the other hand to choose an interpolation function adapted to the surface properties, thus needing a prior recognition of the 3-D structures appearing in the scene. A 3-D region segmentation and interpretation process is thus needed.

2.1.3. Roof Recognition and Reconstruction by Cooperative Combination of Monocular and Binocular Analysis

Building recognition from a single 2-D monocular image analysis is a very difficult task and often leads to an incomplete and sparse set of identified buildings. Building recognition is usually carried out either by global form recognition of regions obtained based upon regions given by low-level segmentation techniques or by perceptual feature (most of the time line segments) grouping techniques. The completeness and the precision of the extracted buildings is very dependent on building complexity, building limit image contrasts, and image quality (MTF, SNR, etc.). Building height computation often relies on building shadow measurements or building vertical edge measurements. Whenever the scene latitude is low and when the image time acquisition is close to noon, building shadow sizes are small and the processed building heights are not precise enough. Moreover, when the buildings are close to the image nadir, their vertical edges are short, thus leading to imprecise height measurements. Thus, an operational urban mapping system cannot rely only on monocular analysis. Most of building mapping techniques dealing with binocular analysis use only stereo information to retrieve elevations of building roofs provided by single monocular analysis [13].

TABLE 3
Theme Exhaustiveness

Themes	Exhaustiveness (samples)	
	Reference DSM	Mapped DSM
All	429000	303045 (70.6%)
Ground	357706	253685 (70.9%)
Roofs	64845	48569 (74.9%)

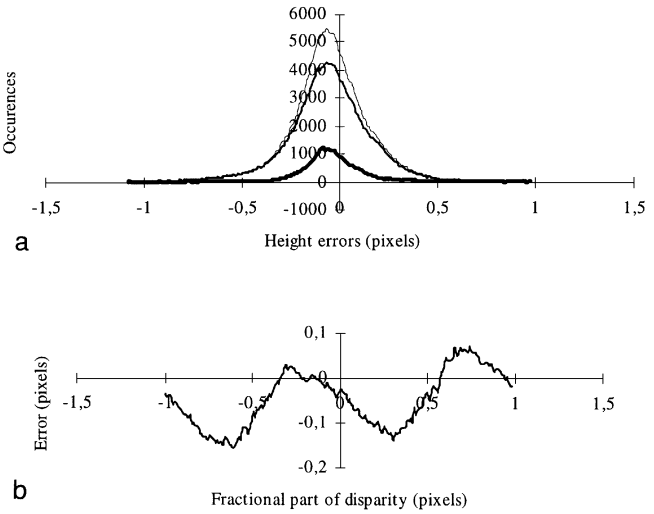


FIG. 11. Matching noise's statistics. (a) Error histogram: roofs (heavy lines), ground (normal line), and all (thin line). (b) Matching noise.

Therefore, we propose to use image and stereo data to carry out both building recognition and building height reconstruction. Our scheme is based on a particular stereo hypothesis. We have restricted our study to the stereo configurations in which building matched pixels can only belong to roof structures and not building vertical faces. The building recognition scheme then becomes a roof recognition scheme. The image data will only be used for retrieving and modeling building limits. The use of range data [2, 18] in image analysis, if its quality is dense and reliable enough, eases the recognition process and greatly eases feature grouping. The roof detection and reconstruction techniques we describe here strongly rely upon the quality of the DSM previously computed by mask template matching and especially on the fact that the inner building range data spreads over nearly all the roof surface and most of the time spreads up to the roofs' limits. Those particular properties will be very helpful for 3-D raised structure extraction and 2-D contour grouping of roof limits. Our roof recognition and building reconstruction scheme can be summarized in Fig. 12.

2.1.3.1. 2-D feature filtering. Feature grouping techniques are very sensitive to contour density. We aim here at reducing the set of initial 2-D contours used to constrain the adaptive

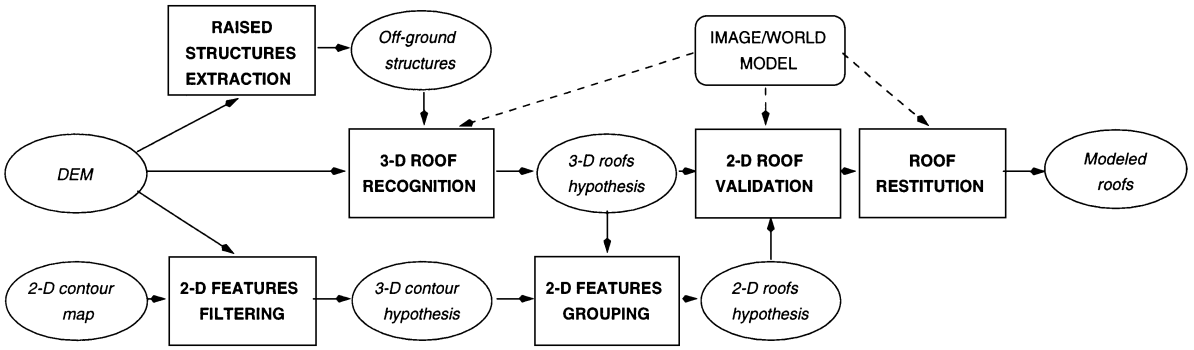


FIG. 12. Building reconstruction and detection.

template matching to a restrained set of 2-D interest contour points corresponding to hypothetical 3-D discontinuity contour points, by studying the local slope around every 2-D contour point. A 2-D contour pixel will be transformed into a 3-D contour hypothesis if the absolute difference between maximum and minimum elevation in a 5×5 neighborhood, is greater than a 1-m threshold (3 times the elevation noise standard deviation). Figures 13a (before filtering) and 13b (after filtering) show the impressive results of the filtering.

All discarded pixels, mainly those corresponding to texture contours, are then filled up in the DSM (disparities were not computed for contour points) with a median of neighbors, thus easing the following the 3-D region extraction.

2.1.3.2. Raised structure extraction. The DSM samples describe the sensor's visible surface. They thus belong to ground structures or raised structures. The raised structure extraction is based first on an altimetric region extraction. Figure 13c shows the 3-D regions computed by a 4-connected region grower on the filled up DSM. The region grower merging criterion is based on a local elevation difference. A pixel is merged to a region if the absolute difference between its elevation and that of the neighboring pixel already belonging to the region is smaller than a threshold. This threshold is fixed to 1 m, which corresponds to 3 times the elevation noise standard deviation (estimated by the simulation-based quality evaluation of the DSM) assuming that elevation noise follows a Gaussian distribution.

The identification of raised structures is based on the analysis of the elevation disparities around the regions' frontiers. An elevation region is considered a raised structure if the altitude in the close inner side of the region's frontier is superior to the altitude on the close outer side of the region on a major part of the region frontier. The results shown in Fig. 13d have been obtained with a height threshold fixed to 2 m on 90% of the region border. One important roof has been missed because of its relatively low elevation (1 m) above the ground. This raised structure extraction algorithm is valid for simple suburban building structures but would be inefficient in dense urban images with complex aggregate roof structures. Baillard *et al*. propose a raised structure extraction technique coping with these complex buildings [2].

2.1.3.3. 3-D roof recognition and 2-D features grouping. We have limited our building roof recognition study to commonly spread buildings—buildings with right angles and planar roofs. Roof recognition is based on a sorting of all the raised structures. The roof hypotheses are built on a 3-D analysis of the DSM height distribution. A raised structure is considered as a roof hypothesis if a planar model [36], computed by a bilinear square approximation, fits the DSM elevation distribution well enough.

Once raw 3-D roof hypotheses are identified, a feature grouping of all the 3-D contour hypotheses immediately around the roof regions is realized. All connected sets of 3-D contour pixels are linked into curves. For each linked pixel a subpixel localization is computed by searching a maximum value along the gradient direction in the gradient edge detector image.

2.1.3.4. Model-driven roof validation and reconstruction. The 3-D contour curves are then approximated by piecewise linear segments using a recursive boundary splitting polygonal representation to reduce contour localization noise. The localization noise standard deviation for an average linear edge contrast with our image quality and for our Canny–Deriche edge detector ($\alpha = 3.0$) has been estimated at around 0.1 pixels. Thus, our boundary splitting threshold is fixed at 0.3 pixels (3 times the localization noise standard deviation). However, the segments defined by the computed polygonal vertices are not optimally located (due to the fact that the vertices belong to the boundary). Therefore, an optimal line segment localization is obtained by a least-squares approximation of all the contour points between two consecutive polygonal vertices. The line segments can be sorted into two classes:

- the first class groups all the segments forming an angle close to the longer segment in object space with a 5° tolerance;
- the second class groups all segments forming a near right angle with the longer segment in object space with a 5° tolerance.

Any segment not belonging to one of these classes is rejected. A 3-D roof hypothesis is validated if there is at least one segment of each class at the consecutive line segments along the roof perimeter.

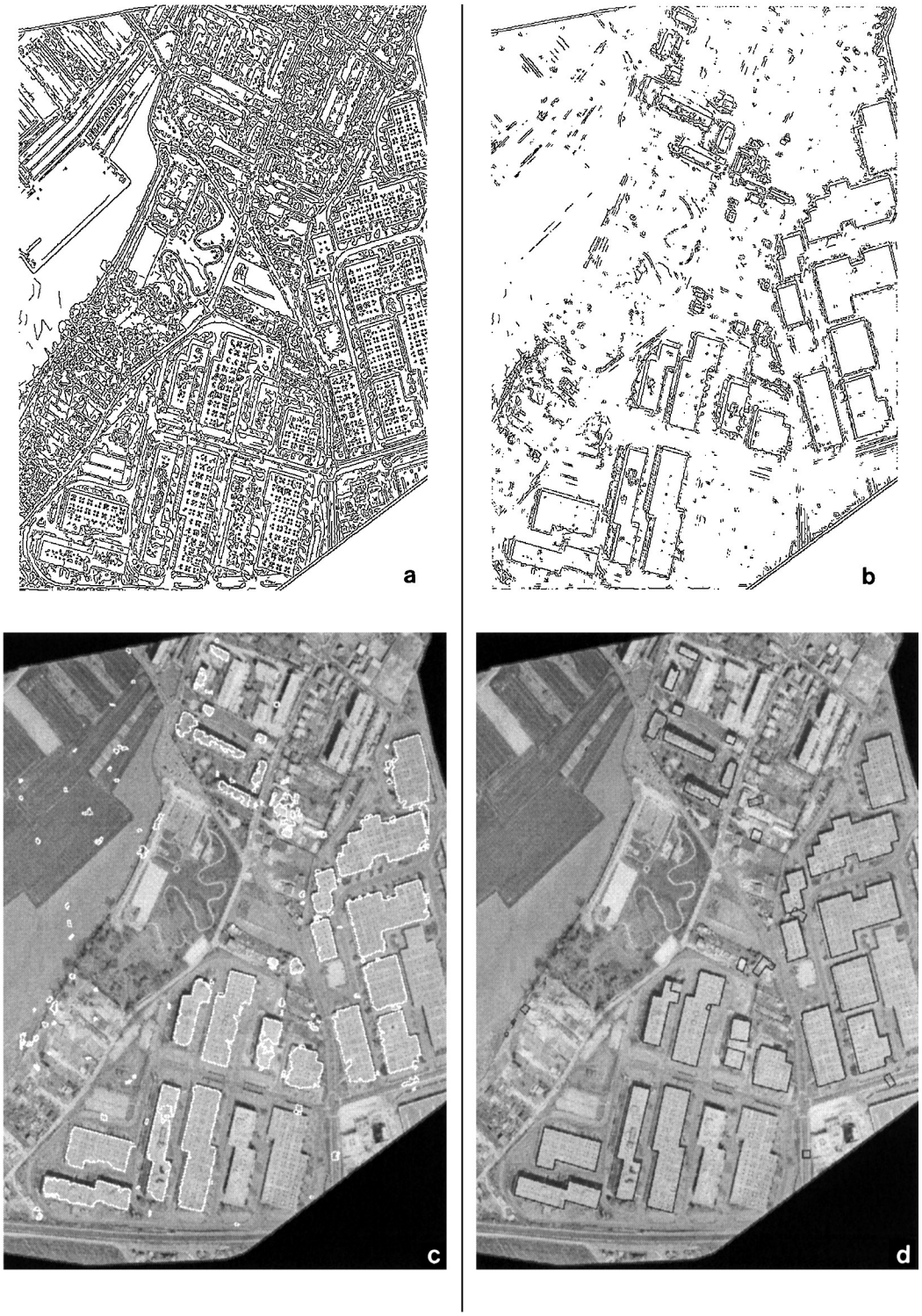


FIG. 13. 2-D feature filtering and raised structures extraction: (a) original image contours, (b) filtered 3-D contours, (c) detected raised structures, and (d) modeled roof limits.

Once raw roof forms have been identified, a model-driven building reconstruction based on the previously computed line segments aims at recovering closed shapes. The line segment grouping follows the following two rules:

- if any two or more consecutive line segments have close orientations and are lined up, they are merged into a new line segment given by the piecewise approximation of all pixels which gave birth to each segment description;
- if any two consecutive line segments form an orthogonal L-junction in object space, the junction point S is kept and it defines a building roof corner.

Some roof forms remain open if two consecutive line segments have close orientations and are not lined up. This is due to a missing feature between line segments. This missing feature is replaced by a virtual segment V forming a right angle in object space with the two framing line segments. The localization of the virtual segment is obtained by moving it along the normal direction to its orientation and adapting its position to the best image contrast. The best image contrast is given by the optimization of the C criterion

$$C = \sum_{(i,j)\in\mathcal{I}_r(V)} G(i, j), \tag{4}$$

where $G(i, j) = \sqrt{G_x(i, j)^2 + G_y(i, j)^2}$ is the image gradient for pixel (i, j) and $\mathcal{I}_r(V)$ represents all the pixels describing the virtual segment V.

The computed building roof edges are shown in Fig. 13d. The few remaining building localization flaws are mainly due to local irregular density of the matching process. Thus, DSM computation can be further improved.

The altitudes $S(Z)$ of all the roof corners are obtained by applying the polynomial bilinear roof surface model to all subpixel roof corners image coordinates $S(i, j)$. Their planimetric coordinates $S(X, Y)$ are given by the intersection of the object plane of altitude Z with the optical axis corresponding to pixel (i, j).

2.1.3.5. 3-D building and scene surface reconstruction. The previously computed DSM describes the visible surface of the 3-D world. The terrain can be derived by the elimination of all structures corresponding to "terrain noise": trees, buildings, cars, etc. The "bare-earth" samples are obtained by the removal of all DSM samples corresponding to raised structures. Meanwhile, we do not know how to recover the terrain topography under raised structures of large coverage (e.g., forest canopy). The only "bare-earth" model we can extrapolate is the one under a small expanse of raised structures like cars or buildings assuming that the terrain surface is locally planar. Our "bare-earth" surface model is then modeled as a TIN using a Delaunay triangulation technique on the DSM remaining samples.

Our building world model is restricted to the buildings which can be modeled by a vertical face polyhedral volume. Each building volume is built by joining a roof, a ground foot print, and a set of vertical faces. The building ground foot and the vertical faces are recovered by extruding the roofs onto the ground. All adjacent and colinear vertical faces are merged into a unique vertical face. All building faces are finally oriented in such a way that the vectors normal to the faces run out the building volume. Figure 14a shows the computed building models for a set of industrial buildings and apartment houses.

Scene modeling consists of managing the relationship between both building and terrain 3-D geometries in a coherent way and merging the building models and the terrain model with a coherent management of the building foot print. Figure 14b shows a shaded view of our computed scene model.

2.2. Approach 2: Stereovision Based on Projective Geometry

The aim of the work carried out by Zahid *et al.* [50] was to localize and to rebuild buildings from aerial images of urban scenes. The difference with the previous approach principally relies on the fact that, prior to the matching step, a monocular segmentation and interpretation of the "west" image is carried out in order to recognize building structures [9]. Matching is then restricted to building areas. In this matching approach, images are not registered and buildings are presumed to be polyhedral.

2.2.1. Image Segmentation

The segmentation of the "west" image is performed by a contour extraction and a contour closing, leading to an adjacency graph of regions [8]. A polygonal approximation of the region boundaries is performed. We then check that a parallelogram fits the polygonal approximation [9]. If not, the region is split into several parallelograms, hence new regions are created in the adjacency graph.

2.2.2. Image Interpretation

Each region R is characterized by a feature vector $\mathbf{A}$ with 4 components: area, average gray level, quality rate of the approximation of the region boundary by a parallelogram (shape factor), and compactness (or circularity). Six classes are possible: lawn, grove, building, tree, shadow, and vertical face. The region classification is carried out on the feature vectors.

Our image understanding system is composed in a neural network classifier and a constraint propagation system.

2.2.2.1. Initial labeling by a neural network. A multilayer perceptron is well suited to achieve the initial labeling, the learning is performed by the Gradient Back Propagation algorithm [9]. A solution to sizing the neural network can be proposed here: the one we used is composed of four layers, with two hidden layers. The input layer is made of 4 neurons, each of them being associated to a region feature. The output layer contains as much neurons as classes.

The first hidden layer acts as thresholding the input values, it receives the input vector in the following form: $\mathbf{E} = [e_1, \ldots, e_i, \ldots, e_N]^t$. For a given region R, the components e_i are the normalized analogic components a_i of the vector $\mathbf{A} = [a_1, \ldots, a_i,$

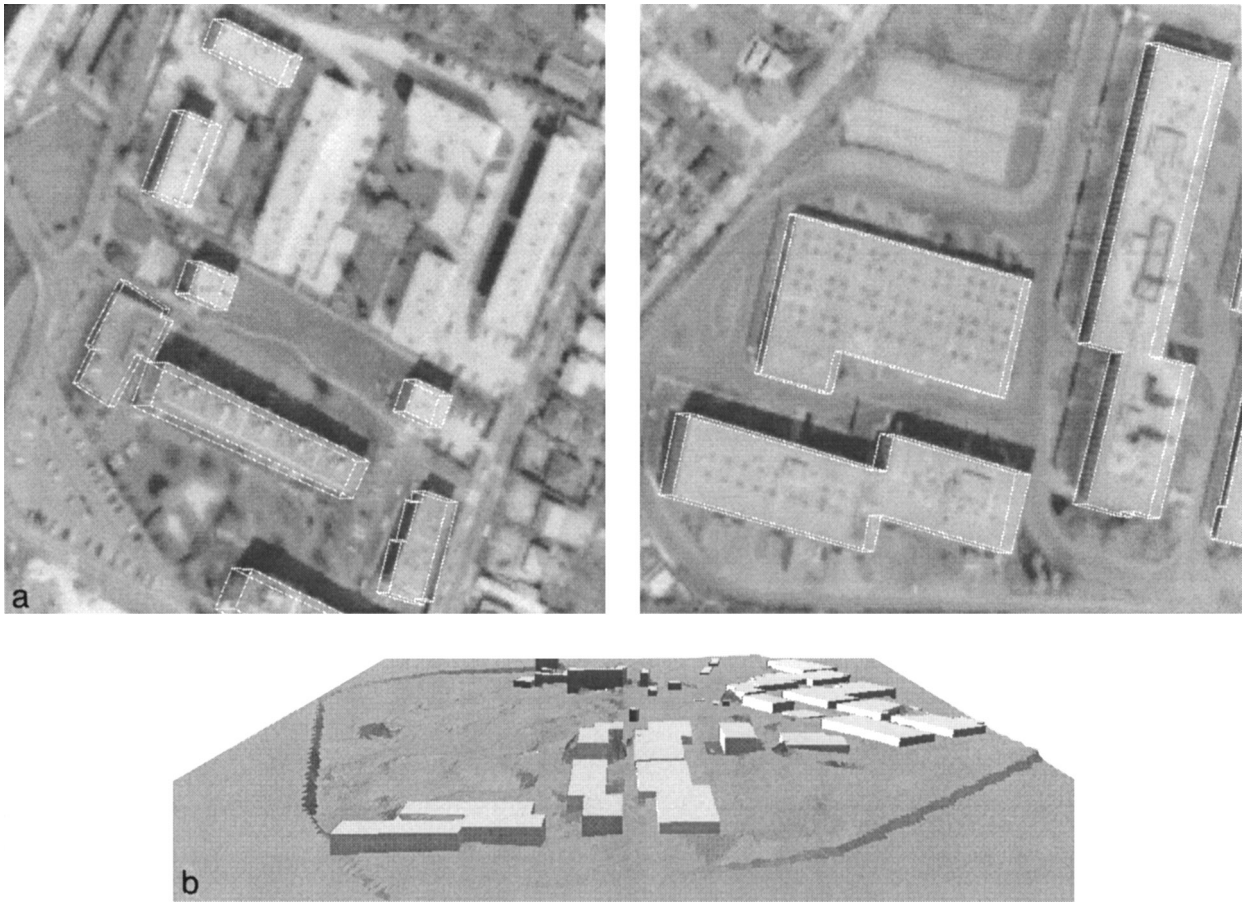


FIG. 14. Building detection and reconstruction. (a) Apartment houses and industrial building reconstruction, (b) shaded view of the 3-D model.

$\ldots, a_M]^t$. In a previous work related in [9], we proposed a method of sizing the layers of the neural network: each neuron of the input layer is connected to the first hidden layer and 3 neurons of the second hidden layer are connected to each output neuron. Hence, the global architecture is then a 4–16–18–6 network.

Let us now explain the initial labeling. The learning set is composed of examples: $\mathbf{C^e} = \{(\mathbf{E_1}, \mathbf{S_1}), \ldots, (\mathbf{E_h}, \mathbf{S_h}), \ldots, (\mathbf{E_L}, \mathbf{S_L})\}$. For a region R_h, $\mathbf{E_h}$ is the vector $\mathbf{A_h}$ normalized and $\mathbf{S_h}$ is the output vector. If $\mathbf{S_h}$ corresponds to a region labeled by the expert, it owns a single $+1$ component, if not, all its components equal -1.

The problem we met in the learning phase was the multiple credit assignment problem:

- two regions with different labels may have almost the same input vector (leading to the same output vector on the first hidden layer).
- a nonlabeled region may have almost the same input vector as a labeled region.

These phenomena lead to a nonconvergence of the network or to a convergence towards a nonsignifiant state; the weights are not properly computed because of the contradictory examples, and the system may not even label a region that had a label in the learning set! To solve this problem, we modified the learning set as follows:

- when two regions R_h and R_p produce two examples $(\mathbf{E_h}, \mathbf{S_h})$ and $(\mathbf{E_p}, \mathbf{S_p})$ with the same binary vector $(\mathbf{H_h} = \mathbf{H_p})$ as ouput of the first hidden layer, both examples are kept but the output vectors $\mathbf{S_h}$ and $\mathbf{S_p}$ are replaced by a new vector:

$$\mathbf{S_h^f} = \mathbf{S_p^f} = \mathbf{S_h} \vee \mathbf{S_p} \tag{5}$$

(f means "fusion" of the vectors) obtained by achieving a logical OR for every component of the output vector (-1 is treated as a logical 0 for this operation).

- this process is applied iteratively to all regions, leading to a new learning set with multiple labels.

The new learning set includes, by principle, identical examples. This is a way to give more importance to certain configurations during the learning.

After this phase of fusion, there are still some examples whose output vectors have all their components to -1. These examples are suppressed of the learning set, because they may have different input vectors, although the same output vector (corresponding to the "unknown" class). Another reason is that the criteria

to not label a region are not always the same, and at least these examples are not numerous.

In conclusion, the learning is thus performed in two stages:

- in a first step, the learning set is $\mathbf{C^e}$; after this first learning, the examples giving the same outputs of the first hidden layer are merged as in relation (5);
- in a second step, the learning is performed with the learning set that contains multiple labels: $\mathbf{C^{fe}} = \{(\mathbf{E_1}, \mathbf{S_1^f}), \ldots, (\mathbf{E_h}, \mathbf{S_h^f}), \ldots, (\mathbf{E_K}, \mathbf{S_K^f})\}$.

2.2.2.2. Final labeling. At the end of the initial labeling, each region is affected with a set of possible labels. The problem is to refine this labeling in order to affect a single label to each region, consistent with the labels of the adjacent regions. This is the role of the constraint propagation system. Constraint propagation is achieved by discrete relaxation [40], which takes into account neighborhood constraints dictated by our knowledge of the scene. For example, a tree may not be included in a building, a shadow must be adjacent to a structure high enough (building, tree, grove, or nonlabeled region), with constraints of relative size, position, and border length.

2.2.3. Matching and 3-D Reconstruction

For all regions labeled "building," a classical correlation technique is used to search a point P' in the "east" image of a corresponding corner P of the "west" image along the epipolar line of P. The template associated to each corner P is not centered at point P but in such a way that the template is completely enclosed in the region.

Refinement of the matching is performed by using the following theorem of the projective geometry. If four points belonging to a quadrilateral could be matched, then it is possible to compute a mapping homography T_r which will give the corresponding point of each "west" point in the "east" image. As one of the image planes is considered as a projective plane, the size of the mapping matrix is 3×3. Its coefficients can be calculated from four pairs of corresponding points expressed in homogeneous coordinates. Matching is performed as follows:

- Given $(P_1, P_2, \ldots, P_n)$ the n corners of the boundary of a region in the "west" image and $(P'_1, P'_2, \ldots, P'_n)$ the corresponding points in the "east" image obtained with the correlation method.
- From these n points and their homologues, the mapping T_r is computed using Widrow–Hoff rules.
- The distance d_i between P_i and $T_r(P_i)$ and the mean distance d_m is calculated.
- For all points with distance d_i greater than d_m, the corresponding point is searched using the correlation with a greater size of window, in order to extend the size of the area in which a corresponding point is searched. Moreover, the sensitivity to noise is reduced.
- The process is iterated until the distance d_m remains stable or until the size of the correlation window is greater than a given threshold.

The matching process requires two thresholds. The first threshold is the size of the strip centered on the current epipolar line to find a corresponding point. Its value results from a statistical study of a set of pairs of corresponding points obtained on test images. The typical value is 2 pixels. The second threshold is the maximum size of the correlation windows. The initial value is fixed to 5×5 pixels. By analyzing results obtained from a set of test images, we observed that the accuracy was almost constant for window size greater than 17×17. The mean distance d_m converges toward 2 pixels.

2.2.4. Results for Approach 2

Figure 15 shows results obtained in two subimages with the method described in Section 2.2; the left column shows original reference images, the middle column shows buildings detected by the interpretation step, and the right column shows perspective views of the modeled buildings obtained by extruding the 3-D roofs onto a virtual ground.

These results show some deficiencies in building detection, due to a bad segmentation and interpretation of the image (some buildings were not recognized by the interpretation process). On the contrary, there is no overdetection and the 3-D relative coordinates of buildings are correct.

3. WORKING ON URBAN AREAS: HIGH-RESOLUTION IMAGERY

Our work is currently moving toward higher resolution imagery, which presents new scope in survey. For higher resolutions, the 3D study of smaller buildings as suburban houses or smaller building structures as chimney tops can be carried out. Only a few recent works deal with high-resolution aerial imagery analysis [15, 20]. This section deals with an extension of our mid-resolution methods to dense urban high-resolution aerial images (8 cm ground resolution).

The content of these images is very different compared to mid-resolution. For example, there are many details which were invisible up toward the occlusions, and the variations of the parallax are very large. To carry out a good surface reconstruction, as decribed in Section 2.1.2.1, both efficient monocular segmentation and matching are required. It is for this reason that, first, we introduce a new edge detector for the segmentation and, second, we develop a specific matching strategy. Finally, some results of surface reconstruction are provided.

3.1. Monocular Segmentation

The segmentation described in Section 2.1.2.1, based on edge detection, must be modified to take into account the characteristics of the high-resolution images.

In these images, the visible vertical sides of all the buildings are very large; they can be considered as significant image areas. The contrast between these areas and roof areas is often very low.

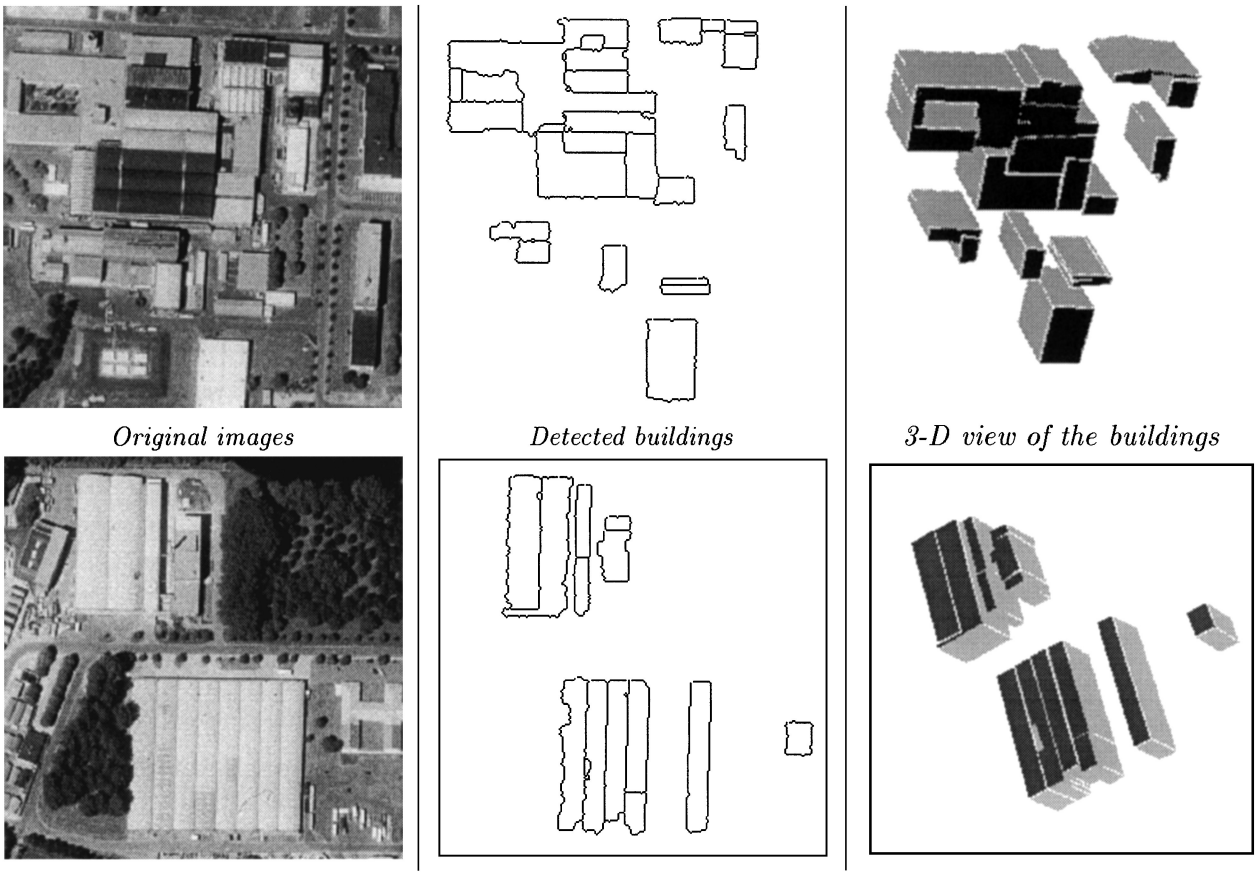


FIG. 15. Building reconstruction using projective geometry.

Furthermore, the areas surrounding the roofs are typically not very homogeneous; the introduction of this detailed structure disturbs the contrast on the edges of the roofs. Moreover, the interior of the roof is not homogeneous; there are always fine structures such as the chimneys or edges which strongly contrast with the other parts of the roof. Tests using edge detectors have shown the persistency of edges resulting from fine structures, while at the same time, whole edges are not detected. Contours are quite often broken and irregular.

That is why a specialized edge detector, introduced in [11], is used. First, a linear derivator filter is adjusted thanks to limit values in order to detect any close contours without error localization. Second, a special thresholding, sorting out the contours by way of geometric and topological attributes of the edges, is used. Low-contrast structures are kept and noise is efficiently filtered. The image segmentation is then obtained by contour closing.

3.2. High-Resolution Image Matching

Both mid-resolution approaches, discussed in Section 2, take advantage of contours. For the high-resolution images the choice of the matching method is discussed according to monocular image characteristics and binocular considerations.

3.2.1. Choice of the Primitive

Occluded area can be very large. Many contours are therefore present only in a single view. Furthermore, geometric deformations occur between both views; segments belonging to vertical sides have a different orientation in the east and west images.

This is why we prefer not to use contour primitives for the stereo matching. In any case, contour primitives or more structured primitives could not be used here without very elaborate matching strategies [21, 31].

A specialized indicator for stereovision [7, 33] shows that our images are extremely textured. It justifies the use of a matching primitive such as the gradient primitive based on the consistency of the local radiometric variation information. We have selected template matching on image gradient vectors with a specific similarity function. Let $(G_E(i, j), G_W(i, j))$ be the respective gradient vector at the point (i, j) of the "east" (respectively, "west") image; on rectangular $(2u+1) \times (2v+1)$ windows, the similarity function is defined by

$$C\,grad((i, j), \xi) = \frac{\sum_{\iota=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} \|G_W(\iota, \eta) - G_E(\iota, \eta+\xi)\|}{\sum_{\iota=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} [\|G_W(\iota, \eta)\| + \|G_E(\iota, \eta+\xi)\|]}. \quad (6)$$

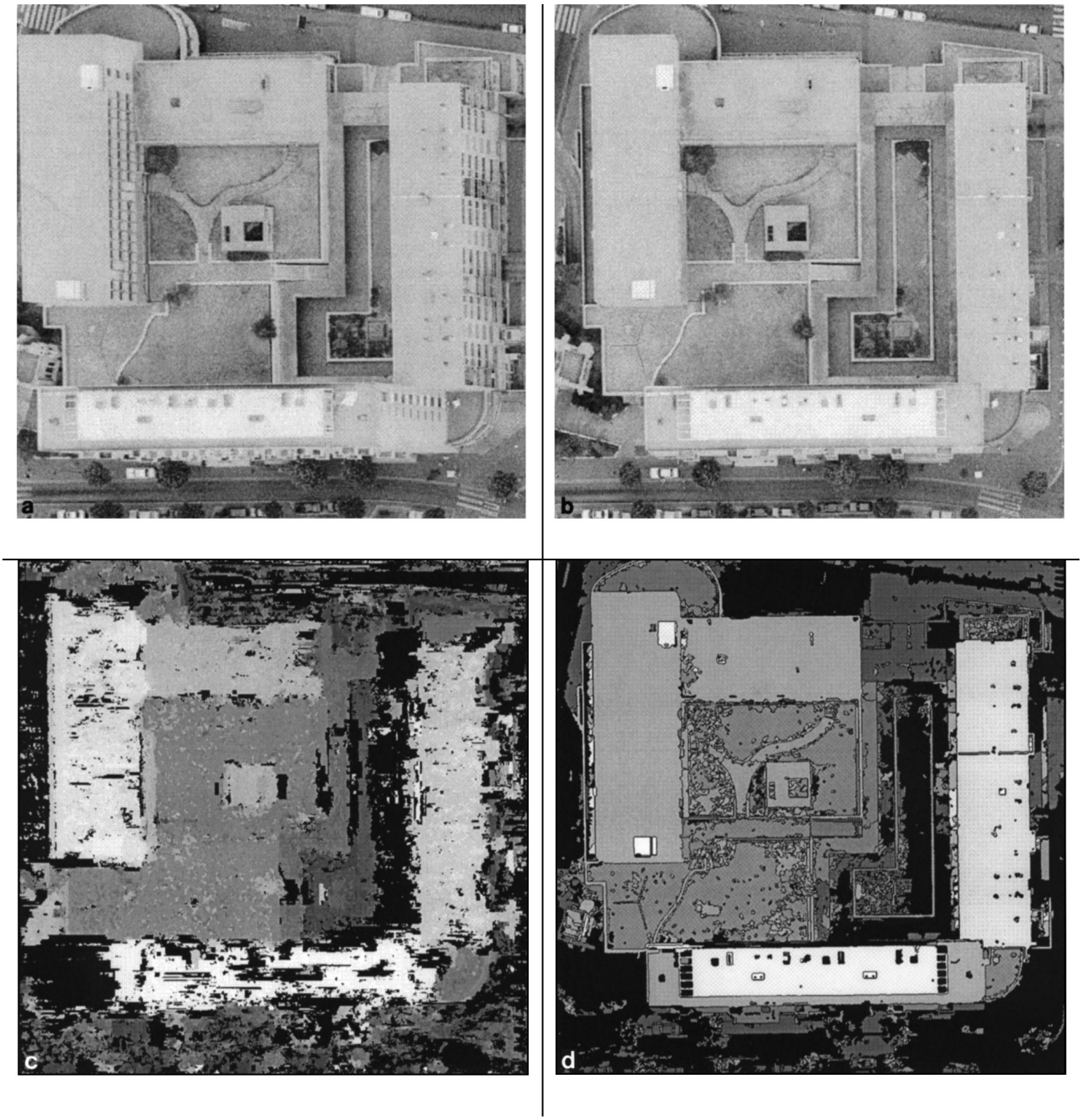


FIG. 16. High-resolution image stereo analysis: (a) east view, (b) west view, (c) altitudes map, and (d) refined altitudes map.

Stereo matching on gradient vectors is very interesting because of the explicit use of the spatial distribution of intensity [42]. However, matching using a single vector is not reliable. The originality of the proposed similarity function is to enable matching on windows of gradients [12]. According to the authors, such a correlation is less sensitive to the noise and the maximum of the correlation curve stands out better than when the cross-correlation function (Eq. (1)) is used.

3.2.2. Multiscale Approach

In these high-resolution images, the variation due to parallax is frequently about 100 pixels, sometimes reaching 200 pixels for subimages (1000×1000) between a point on the floor and a point on the top of a building. This variation in the parallax is so large that it has encouraged us to build a multiscale matching approach with a coarse-to-fine strategy [49].

To build the pyramid of gradients, we use a method similar to the one described in [1]. First, the initial image is smoothed and subsampled to obtain the reduced images. Second, a gradient detector is applied to these images to obtain the multiscale gradient representation.

We selected a filter for the image pyramid construction not only according to the quality measurement of the pyramid but also to that of the whole matching process. Three different filters were tested: an average filter, a median filter, and the Chehikian filter [5], which is a variant of the Burt filter [4]. Results were analyzed on different roof homogeneous areas. Inside each area, the percentage of matched points and statistics on the values of parallax were computed. Results showed that the average and median filters performed equally well and that the more sophisticated Chehikian filter did not provide significantly better results. It is not suited for stereo matching, as the smoothing is too strong between two consecutive levels of the pyramid.

For the gradient pyramid construction, a Deriche filter is applied to compute the gradient vectors for each level of the pyramid.

3.3. Results: Surface Reconstruction

The results are presented on a subimage (980×980) from a $6000 \times 10{,}000$ pixels aerial stereopair (8 cm ground resolution) supplied by the I.G.N. and covering the city of Colombes, France (Fig. 16). The scene is composed of a multiroof complex building with different heights (Figs. 16a and 16b). The disparity map is computed from a 4-level pyramid and from an initial parallax variation of about 180 pixels. The coarse-to-fine strategy is applied. For each level of the pyramid, the matching is obtained from the gradient correlation scheme. Finally, a two-way filtering technique is performed (cf Section 2.1.2.2).

This disparity map is improved using surface modeling. A radiometric segmentation (as described in Section 3.1) is carried out. Then, using the disparity map, we compute for each region (cf Section 2.1.2.1) a mean 3-D planar model (Fig. 16d).

4. CONCLUSION AND FUTURE WORK

This paper has described a set of techniques to perform 3-D stereo mapping on aerial imagery. After performing a survey of computational stereo framework, several complementary stereo techniques are discussed for suburban mid-resolution images. In particular, a hierarchical correlation process and a new edge preserving mask correlation are proposed. Our goal in stereo development has not been just to obtain a DSM, but to compute a complete framework for 3-D building and 3-D scene model reconstruction.

A simulator-based validation environment for optical stereomapping has been described. This approach is unavoidable due to the lack of suitable ground truth data. It allows us to identify and measure the intrinsic algorithmic errors as well as measure the influences of image characteristics on the mapping results. Important applications can been pointed out, namely the parametrical quality assessment and improvements of existing or new algorithms and contribution to the specifications of future middle- or high-resolution optical systems. This simulation-based approach can be applied to any sensor (SAR, radar altimetry, LIDAR) and to any other exploitation technique (multisensor data fusion, dynamic mapping) provided that images can always be simulated in a realistic way.

We believe that our mid-resolution suburban scene reconstruction framework is an improvement to the state of the art. We attribute this success to the cooperative combination of 2-D monocular and 3-D stereo analysis. Monocular analysis is used both for constraining stereo matching by edge feature detection and for retrieving and modeling building limits. Stereo data are used to support feature grouping, building recognition, and building height reconstruction.

However, our analysis has also pointed out that applying the reconstruction techniques that we first investigated for mid-resolution images to high-resolution images is insufficient without careful adaptations. We have thus developed a new matching technique based on a multiscale gradient correlation method to improve the density and the reliability of the matching process. There are several fields of investigation for our future work. Our forthcoming work will first integrate different complementary stereo techniques (adaptive window form matching, adaptive window size matching, and multiscale correlation) in a unified matching method in object space. Then, our roof recognition and building reconstruction scheme will be adapted to the modeling of more complex buildings such as aggregated multiroof structures.

Moreover, the use of color images would be very helpful both for image analysis (better structure discrimination) and for easier pattern recognition. Some recent results lead us to believe that the use of more than 8-bit gray-scale digital images will definitely be a major improvement for stereo matching by reducing the number of mismatches for homogeneous roofs and especially for dark shadows. Major problems encountered in scanned imagery due to unsatisfying image quality and to nonlinearity of photosensitivity will be thus resolved.

ACKNOWLEDGMENTS

Part of this work was supported by *Aérospatiale Espace & Défense, 100, bd. du Midi, 06156 Cannes-la-Bocca Cédex, France*. Another part stems from a research contract between ETIS and the *Institut Géographique National*, which supplied the aerial images used in this paper. We also thank the reviewers for their very helpful comments.

REFERENCES

1. R. Azencott, F. Durbin, and J. Paumard, Robust recognition of buildings in compressed large aerial scenes, in *Int. Conf. on Image Processing, Lausanne, 1996*, pp. 617–620.
2. C. Baillard, O. Dissard, O. Jamet, and H. Maître, Extraction and characterization of above-ground areas in a peri-urban context, in *Proc. of the IAPR/TC-7 workshop on methods for extracting buildings, roads and other man-made structures from images, Graz, 1996*.

3. S. T. Barnard and M. A. Fischler, Computational stereo, *ACM Comput. Survey* **14**(4), 1982, 553–572.
4. P. J. Burt, The Laplacian Pyramid as a compact image code, *IEEE Trans. Commun*. **31**(4), 1984, 532–540.
5. A. Chehikian, Algorithmes optimaux pour la génération de pyramides d'images passe-bas et laplaciennes, *Traitement du Signal* **9**(4), 1992, 297–307.
6. R. Chung and R. Nevatia, Use of monocular groupings and occlusion analysis in a hierarchical stereo system, *Comput. Vision Image Understanding* **62**(3), 1995, 245–268.
7. S. D. Cochran and G. Medioni, 3-D surface description from binocular stereo, *IEEE Trans. Pattern Anal. Machine Intell*. **14**(10), 1992, 981–994.
8. J.-P. Cocquerez, R. Deriche, and G. Almouzni, An efficient method to build early image description, in *9th Int. Conf. on Pattern Recognition, Rome, 1988*.
9. J.-P. Cocquerez, Ph. Gaussier, S. Philipp, and M. Zahid, Expert system and connexionist networks for complex aerial picture analysis, in *Proc. of 7th Scandinavian Conference on Image Analysis, Aalborg (DK), 1991*.
10. R. T. Collins, Y. Q. Cheng, C. Jaynes, F. Stolle, X. Wang, A. R. Hanson, and E. M. Riseman, Site model acquisition and extension from aerial images, in *Int. Conference on Computer Vision, Cambridge, MA, 1995*.
11. M. Cord, F. Huet, and S. Philipp, Optimal adjusting of edge detectors to extract close contours, in *Proc. of 10th Scandinavian Conference on Image Analysis, Lappeenranta, Finland, June 1997*.
12. A. Crouzil, L. Massip-Pailhes, and S. Castan, Mise en correspondance par corrélation des gradients, in *Reconnaissance des formes et intelligence artificielle, Rennes, Jan*. 1996, *AFCET*, pp. 236–245.
13. T. Dang, O. Jamet, and H. Maître, Applying perceptual grouping and surface models to the detection and stereo reconstruction of building in aerial imagery, in *Proc. of ISPRS Commission III Symposium, Munich, 1994*, pp. 165–172.
14. R. Deriche, Using Canny's criteria to derive a recursively implemented optimal edge detector, *Int. J. Comput. Vision* **1**(2), 1987, 167–187.
15. D. Frère, M. Hendrickx, J. Vandekerckhove, T. Moons, and L. Van Gool, On the reconstruction of urban house roofs from aerial images, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images II* (A. Grün, E. Baltsavias, and O. Henricsson, Eds.), pp. 87–96, Birkhäuser Verlag, Basel, 1997.
16. A. Grün, E. P. Baltsavias, and O. Henricsson (Eds.), *Automatic Extraction of Man-Made Objects from Aerial and Space Images (II)*, Birkhäuser Verlag, Basel, 1997.
17. A. Grün, O. Kübler, and P. Aggouris (Eds.), *Automatic Extraction of Man-Made Objects from Aerial and Space Images*, Birkhäuser Verlag, Basel, 1995.
18. N. Haala, Detection of buildings by fusion of range and image data, in *Proc. of ISPRS Commission III Symposium, Munich, 1994*.
19. N. Haala, M. Cramer, and J. Kilian, Sensor fusion for airborne 3D data capture, in *2nd Int. Airborne Remote Sensing Conference, San Francisco, 1996*, pp. 344–353.
20. O. Henricsson and E. Baltsavias, 3-D building reconstruction with ARUBA: A qualitative and quantitative evaluation, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images II* (A. Grün, E. Baltsavias, and O. Henricsson, Eds.), pp. 65–76, Birkhäuser Verlag, Basel, 1997.
21. O. Henricsson, F. Bignone, W. Willuhn, F. Ade, O. Kübler, E. Baltsavias, S. Mason, and A. Grün, Project AMOBE: Strategies, current status and future work, in *Int. Archives of Photogrammetry and Remote Sensing, Wien, Austria, 1996, ISPRS*, pp. 321–330.
22. M. Herman and T. Kanade, Incremental reconstruction of 3-D scenes from multiple, complex images, *Artificial Intell*. **30**, 1986.
23. Y. C. Hsieh, D. M. MacKeown, and F. P. Perlant, Performance evaluation of scene registration and stereo matching for cartographic feature extraction, *IEEE Trans. Pattern Anal. Machine Intell*. **14**(2), 1992, 214–238.
24. R. B. Irvin and D. M. MacKeown, Methods for exploiting the relationships between buildings and their shadows in aerial imagery, *IEEE Trans. Systems Man Cybernet*. **19**(6), 1989.
25. C. O. Jaynes, F. R. Stolle, H. Schultz, R. T. Collins, A. R. Hanson, and E. M. Riseman, Three-dimensional grouping and information fusion for site modeling from aerial images, in *Proc. of Arpa Image Understanding Workshop, Palm Springs, CA, 1996*, pp. 479–490.
26. D. G. Jones, *Computational Models of Binocular Vision*, Ph.D. thesis, Stanford University, 1991.
27. M. Jordan and J.-P. Cocquerez, Three-dimensional description of scenes observed in aerial photography, *Pattern Recognit*. **28**(7), 1995.
28. T. Kanade and M. Okutomi, A stereo matching algorithm with an adaptative window: Theory and experiment, *IEEE Trans. Pattern Anal. Machine Intell*. **16**(9), 1994, 920–932.
29. T. Kim and J.-P. Muller, Automated building height extraction and building detection from high resolution aerial and space imagery, in *IAPR Workshop on Machine Vision Applications, Kawasaki, Dec. 1994*, pp. 364–367.
30. T. Kim and J. P. Muller, Automated building height estimation and object extraction from multiresolution imagery, in *SPIE Conf. on Integrating Photogrammetric Techniques with Scenes Analysis and Machine Vision II, Orlando, FL, April 1995*, pp. 267–276.
31. F. Lang and W. Förstner, Surface reconstruction of man-made objects using polymorphic mid-level features and generic scene knowledge, *Int. Archives Photogram. Remote Sensing* **31**, B3, 1996, 415–420.
32. C. Lin, A. Huertas, and R. Nevatia, Detection of buildings from monocular images, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images* (A. Grün, O. Kübler, and P. Aggouris, Eds.), pp. 125–134, Birkhäuser Verlag, Basel, 1995.
33. J.-L. Lotti, *Mise en correspondance stéréo par fenêtres adaptatives en imagerie aérienne haute résolution*, Ph.D. thesis, Université de Nice, Sophia Antipolis, 1996.
34. J.-L. Lotti and G. Giraudon, Adaptative window algorithm for aerial image stereo, in *12th Int. Conf. on Pattern Recognition, Jerusalem, 1994*, pp. 701–703.
35. D. M. MacKeown, S. D. Cochran, and S. J. Ford, Research in automated analysis of remotely sensed imagery: 1993–1994, in *Proceedings of ARPA Image Understanding Workshop, Monterey, CA, Nov. 1994*, pp. 99–132.
36. H. Maître and W. Luo, Using models to improve stereo reconstruction, *IEEE Trans. Pattern Anal. Machine Intell*. **14**(2), 1992, 269–277.
37. R. Mohan and R. Nevatia, Using perceptual organization to extract 3-D structures, *IEEE Trans. Pattern Anal. Machine Intell*. **11**(11), 1989, 1121–1138.
38. N. Paparoditis, *Reconstruction 3D de scènes urbaines à partir d'images satellites stéréoscopiques haute résolution*, Ph.D. thesis, Univ. de Nice, Sophia Antipolis, 1997.
39. N. Paparoditis, L. Polidori, and E. Savaria, Parametric simulation-based evaluation of 3-D mapping techniques on optical high resolution remote sensing images, in *Proc. of SPIE Europto Conf., Paris, France, 1994*.
40. A. Rosenfeld, R. A. Hummel, and S. W. Zucker, Scene labeling by relaxation operations, *IEEE Trans. Systems, Man, Cybernet*. **6**, 1976, 420–433.
41. M. Roux and D. M. MacKeown, Feature matching for building extraction from multiple views, in *Proceedings of ARPA Image Understanding Workshop, Monterey, CA, Nov. 13–16, 1994*, pp. 331–349.
42. D. Scharstein, Matching images by comparing their gradient fields, in *Proc. 12th IAPR Int. Conf. on Pattern Recognition, Jerusalem, Israel, 1994*, pp. 572–575.
43. T. Schenk, A layered abduction model of building recognition, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images* (A. Grün,

O. Kübler, and P. Aggouris, Eds.), pp. 117–124, Birkhäuser Verlag, Basel, 1995.

44. H. Schultz, Terrain reconstruction from widely separated images, in *Integrating Photogrammetric Techniques with Scene Analyis and Machine Vision II, Orlando, FL, April 1995, SPIE*.
45. J.-P. Thirion, Realistic 3D simulations of shapes and shadows for image processing, *Comput. Vision Graphics Image Process*. **54**(1), 1992, 82–90.
46. U. Weidner, An approach to building extraction from digital surface models, in *Proc. of 18th ISPRS Congress, Vienna, Austria, 1996.*
47. U. Weidner and W. Förstner, Towards automatic building reconstruction from high resolution digital elevation models, *ISPRS J*. **50**(4), 1995, 38–49.
48. H. Wiman and P. Axelsson, Finding 3D-structures in multiple aerial images using lines and regions, *Int. Archives Photogram. Remote Sensing* **31**, B3, 1996, 953–959.
49. A. Witkin, D. Terzopoulos, and M. Kass, Signal matching through scale space, *Int. J. Comput. Vision* 1987, 133–144.
50. M. Zahid, M. Chapron, and J.-P. Cocquerez, 3D reconstruction of buildings using stereovision based on projective geometry, in *Proc. of 7th ICIAP, Bari, Italy*, 1993.