Computer Vision and Image Understanding
Vol. 76, No. 3, December, pp. 244–258, 1999
Article ID cviu.1999.0793, available online at http://www.idealibrary.com on IDEAL®

3-D Reconstruction of Urban Scenes from Aerial Stereo Imagery: A Focusing Strategy

C. Baillard*

Institut Géographique National, Laboratoire MATIS, 2, av. Pasteur, 94160 Saint Mandé, France

and

H. Maître

ENST, Département TSI, 46, rue Barrault, 75013 Paris, France
E-mail: maitre@ima.enst.fr

Received October 19, 1998; accepted June 21, 1999

A contribution to the automatic 3-D reconstruction of complex urban scenes from aerial stereo pairs is proposed. It consists of segmenting the scene into two different kinds of components: the ground and the above-ground objects. The above-ground objects are classified either as buildings or as vegetation. The idea is to define appropriate regions of interest in order to achieve a relevant 3-D reconstruction. For that purpose, a digital elevation model of the scene is first computed and segmented into above-ground regions using a Markov random field model. Then a radiometric analysis is used to classify above-ground regions as building or vegetation, leading to the determination of the final above-ground objects. The originality of the method is its ability to cope with extended above-ground areas, even in case of a sloping ground surface. This characteristic is necessary in a urban environment. Results are very robust to image and scene variability, and they enable the utilization of appropriate local 3-D reconstruction algorithms. © 1999 Academic Press

Key Words: **aerial imagery; urban scenes; above-ground; buildings; vegetation; regions of interest; 3-D data analysis; classification; Markov random field.**

1. INTRODUCTION

Over the past few years, the 3-D reconstruction of urban scenes from images has become a key issue in many applications: cartography, urbanism, simulation, monitoring, etc. Much work has been done on the automatic extraction of 3-D information from urban stereo pairs, including stereo matching, building extraction and reconstruction, or change detection. But urban environments are extremely difficult to handle, for several reasons:

- *complexity*: the 3-D model of the scene is very complex, with many height discontinuities and large differences in height;
- *diversity*: the geometric and thematic diversity of the objects composing the scene requires the use of various 3-D models, with appropriate detail levels.
- *density*: the high density of above-ground objects (buildings and vegetation), often adjacent to each other, leads to many hidden parts, many shadows, and complex aggregates.

Occluded areas, periodic structures (parallel borders of buildings or roads), homogeneous areas (shadows, roofs), and moving objects (cars, trucks) make the stereoscopic matching process often ambiguous. In addition, since large depth discontinuities are frequent, geometric constraints about the surface must be used very carefully. The complexity and the diversity of the objects composing the scene prevent us from using simple 3-D models for the reconstruction. The requirements for the reconstruction itself, i.e., geometric accuracy and detail level, differs significantly according to the objects and to their context.

Due to these difficulties, the automatic 3-D reconstruction of urban scenes has rarely been approached in a global way. Most studies have focused on the local reconstruction of a certain kind of buildings.

The first approaches consisted of detecting buildings from a single image [1–4]. Roof elevations can then be retrieved from shadows [5, 6] or from the lengths of vertical borders [7]. The height of buildings can also be computed by matching structures detected in differents images [8–10] or by using independently produced 3-D points [11, 12]. These methods require the buildings to be isolated with a rectilinear shape.

In the last few years, the analysis of 3-D data such as 3-D lines, 3-D corners, or DEM (digital elevation model) has seen much development. Still these latter approaches have been mostly dedicated to the reconstruction of specific buildings: rectilinear shape [13–16], flat roof [17], or simply parameterizable shape [14, 18–20]. An attempt to automatically select the appropriate model has been done in [21]. Some methods based on coplanar grouping of 3-D lines have been proposed for reconstructing

* Currently working at Oxford University, in the Robotics Research Group (Department of Engineering Science). E-mail: caroline@robots.ox.ac.uk.

1077-3142/99/ $30.00

generic models of buildings from high-resolution multiple imagery [22–24]. However, buildings must still be isolated or localized in advance. In case of a dense urban environment, where most objects are aggregated to each other, some results have been provided completely automatically with high resolution imagery only [25–27].

No satisfying method for an automatic reconstruction from images at a medium resolution (between 50 cm and 1 m per pixel) has been proposed to date. All methods rely on strong assumptions, either about the object geometry or about its close neighborhood. The application of local reconstruction methods appears to be essential to successfully reconstruct urban scenes at a medium resolution, because the number of features is then reduced, and relevant geometric rules about shape can be applied.

Therefore, we believe that a focusing strategy is necessary to handle complex scenes and overcome the classical limits of 3-D reconstruction. The approach presented here takes place within a two-stage reconstruction strategy which only needs two stereo images as input. First a global analysis of the scene involving both radiometry and altimetry produces a coarse but reliable information about the whole scene, and regions of interest (ROI) can be defined. Then the local and contextual information related to these ROI can be used to drive the local application of relevant reconstruction algorithms. The definition and characterization of ROI have many advantages for the reconstruction:

- the selection of relevant features dramatically reduces the number of possible combinations (for grouping methods, for instance), leading to reduced risk of errors,
- the characterization of the ROI can be used to select an appropriate object model or a relevant reconstruction method, in order to get more accurate results,
- the parameterization can be driven by the quantitative information provided by the ROI (initialization of active contours or parametric models, for instance).

If many local reconstruction algorithms have been proposed already, the issue concerning the detection of ROI from urban imagery has rarely been studied. External data can be used, such as maps [28] or 2-D GIS [29, 30]. But additional data are not always available, and their utilization is often difficult. A few methods have been proposed for detecting buildings from a DEM, but only in the case of isolated buildings (see Section 2.2 for a quick review). Given the lack of previous work, our efforts have been dedicated to the detection of ROI in complex urban environments.

More precisely, we have been interested in the segmentation of the scene into above-ground objects, or AGO, consisting of buildings or trees. They are key features of the urban scene, since they structure it and define relevant areas of interest for a local and specific reconstruction. They are also closely related to the digital terrain model (DTM), which describes the ground surface only. In addition, most surface discontinuities and hidden areas in images are due to AGO. Hence, they play a decisive role in the reconstruction process.

This paper is organized as follow: in Section 2 the input data are presented and our strategy for segmenting the scene into above-ground objects is introduced. Section 3 describes the main stages of the process: First a graph is produced from 3-D data; then each node of the graph is classified as *ground* or *above-ground*; finally the radiometric information is introduced to define *building* and *vegetation* objects. Results are shown and discussed in Section 4.

2. A FOCUSING STRATEGY FOR THE 3-D RECONSTRUCTION OF URBAN SCENES

2.1. Input Data and Objectives

The input data are two stereo aerial images, at a medium scale (typically 40 cm per pixel resolution) and panchromatic with 256 gray levels. The epipolar geometry is known, which will be used for stereo matching. The scene can be arbitrarily complex, with no restriction on object density or geometry. Figure 1 shows an example of such a stereo pair, on which the method will be illustrated.

Our reconstruction strategy relies on a preliminary detection and characterization of AGOs, in order to achieve a relevant 3-D reconstruction of the scene. As areas of interest, AGOs do not have to be accurately delineated, but their detection must be exhaustive. Importantly, the model of above-ground must be generic in order to cope with all typical situations in urban scenes. Extended above-ground areas of any shape and size, like wooded areas, large buildings, or blocks of houses, must be entirely detected. Adjacent buildings and trees, or adjacent buildings of different height, must be detected and separated. The process must also cope with sloping ground or roofs. Figures 2 and 3 show typical examples of complex configurations in an urban environment.

2.2. Related Work to Above-Ground Detection

Little work has been done about automatic above-ground detection. It is commonly assumed that the above-ground objects are small or isolated. We have distinguished two classes of methods, both based on DEM analysis.

The first approach consists of subtracting a DTM from the DEM, where the DTM is usually produced by applying a morphological opening to the DEM ("top hat" filtering). This method is commonly used for low-resolution imagery (typically satellite imagery) or for smooth surfaces. But it cannot be applied in an urban environment, because the size of the structuring element is a maximal size for what is detected above ground. As a result, the extended above-ground and urban aggregates (i.e., blocks of buildings and trees), such as those shown in Fig. 2, cannot be detected.

The other approach to above-ground extraction consists of segmenting the DEM into relevant regions. Baltsavias *et al.*

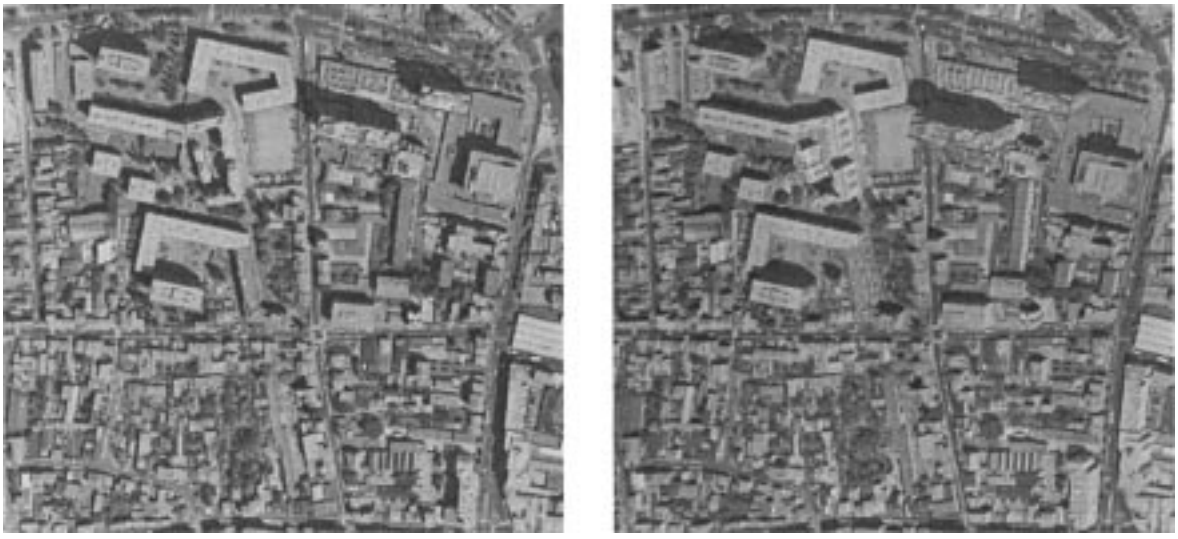

FIG. 1. Aerial stereo pair of the suburb of Paris. The images are 1105 × 1024, with one pixel corresponding to a ground length of 40 cm. One pixel difference in disparity (measured between the two images) corresponds to 96 cm difference in height.

have proposed using a range image edge detector or to group heights into consecutive height ranges [31]. This solution is closely related to studying DEM isolines [32]. An alternative solution focused on relative positions between regions has been proposed in [17]: the surface is segmented into homogeneous regions using a classical segmentation algorithm based on local height discontinuities and then regions are sorted into two groups according to their relative elevation. However, a very reliable DEM is required (a multiview matching algorithm is used) and the classification process is local. It cannot deal with complex situations such as those shown in Fig. 3.

The thematic analysis of above-ground has also often been limited to the selection of isolated buildings according to certain shape and size criteria, sometimes involving radiometric or spectral information [31, 22]. Some attempts have been made to distinguish adjacent buildings and vegetation by using textural filters and a learning stage on radiometry [33] or by using 3-D laser data [34].

2.3. Overview of the Method

The approach presented in this paper relies on a global analysis of the scene consisting of two modules (see Fig. 4). First a DEM is computed from the images by stereo matching. Then a classification stage is performed, which segments the scene into AGO and simultaneously produces a DTM of the scene.

Computation of a DEM. The DEM is computed with the automatic matching algorithm described in [35, 35a]. This algorithm has been especially designed for urban environments. It relies on successive complementary matching steps, all of which are performed by dynamic programming. It takes advantage of both feature-based and area-based matching strategies.

First, intensity edges of both images are matched, which produces piecewise continuous 3-D chains. This provides a description of the scene structure containing the highest elevation of most height discontinuities.

FIG. 2. Two typical examples of above-ground aggregates (details from Fig. 1).

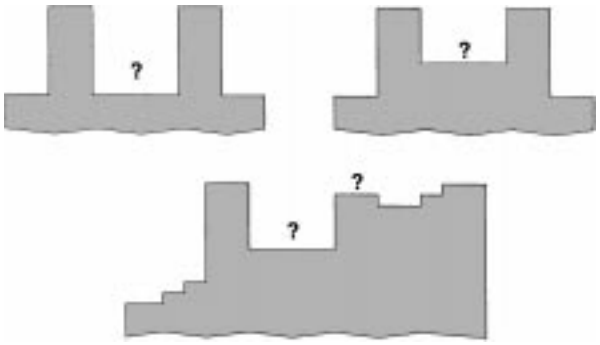

FIG. 3. Examples of ambiguous configurations of DEM (in 2-D), as seen for building silhouettes. A local classification process is not sufficient to separate the above-ground from the ground, particularly for the regions marked '?'.

Then the pairs of intervals defined by the matched edges are matched within a two-step area-based matching process. A strong radiometric similarity constraint is first applied in order to produce only reliable pairs. Then a second area-based matching step is performed with a looser radiometric constraint but a stronger geometric one (smoothness constraint) to complete the 3-D information on unmatched areas. This strategy relies on the assumption that local extrema of depth along epipolar lines are recovered as reliable pairs during the first area-based step.

Each matching step is performed by dynamic programming for each pair of epipolar lines. Dynamic programming is a powerful matching strategy, because it provides an optimal solution for each epipolar line pair, involving all consistency constraints along these lines: unicity and order, but also duality between discontinuity and occlusion. In particular, the hidden parts are not matched. Therefore, local constraints like the disparity range or thresholds on similarity measure can be released without significantly increasing errors. This aspect is especially important in an urban context where differences in height can be very large around towers and the correlation values very low on homogeneous areas.

This hierarchical algorithm has proved reliable (producing few noisy and altimetrically accurate 3-D data), fast, and robust to image variability. Figure 5 shows the DEM derived from the pair of Fig. 1. It is complete and reliable with all surface discontinuities preserved, although slightly delocalized. Importantly, the areas of the scene which are not visible in both images (occlusion areas) are not matched to guarantee the reliability of the 3-D data. As a result, the DEM is not completely dense, since these areas have no altimetric information.

Classification. The classification stage is the main issue of this paper. It relies on the following definitions:

- an *above-ground region* is a part of the scene higher than the ground, from a critical value, in a given neighborhood;
- an *above-ground object* is a monothematic (building or vegetation) above-ground region with locally homogeneous elevations.

Note that these definitions are local and only based on rather low-level information. It is also assumed that the ground surface, represented by the DTM, is continuous with a limited slope.

A block diagram of the classification process is shown in Fig. 6. It uses both radiometry from images and altimetry from the DEM. The first part of the analysis relies on 3-D information. An adjacency graph is derived from the DEM, where a node represents a region of the scene with homogeneous elevation. Then each node of the graph is labeled as ground or above-ground, following a Markovian labeling scheme. Given these labels, radiometric criteria separate building from vegetation nodes. The final above-ground objects (buildings and trees) are produced by merging adjacent nodes according to altimetric and topological criteria. Each step of the process is detailed in the next section.

3. DESCRIPTION OF THE CLASSIFICATION PROCESS

3.1. Creation of a Graph from the DEM

The DEM consists of two kinds of points: those 3-D points derived from the matched pixels and those with unknown height for which no correspondence was found.

The set of 3-D points is segmented into homogeneous altitude regions by a classical growing and merging algorithm. The growing step consists of starting from a point then aggregating neighboring points with close elevation (elevation difference below a threshold δ_g). Each region is then associated with one

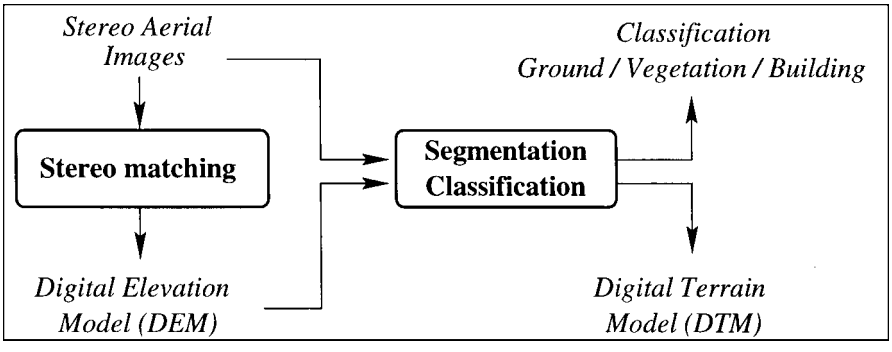

FIG. 4. Global analysis of the scene from a stereo pair: Overview of the method.

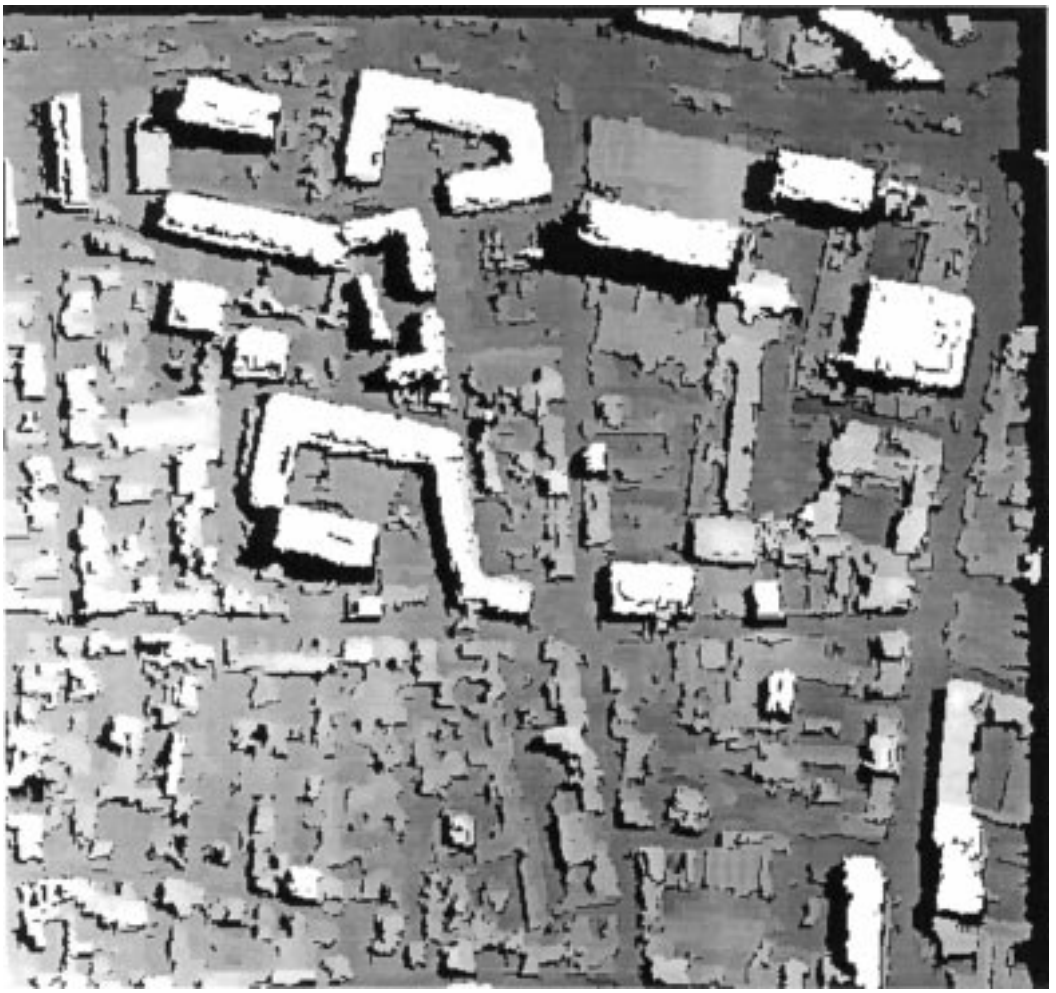

FIG. 5. Digital elevation model automatically computed from the pair of Fig. 1 by dense stereo matching. The elevation data are represented in an orthographic system, without perspective distortion. Black pixels have not been unmatched, most often because they are not visible in one of the two images.

representative elevation value. During the merging step, small regions are merged to the neighboring ones with the closest elevation.

Areas with unknown height elevation points are called *elevation gaps*. The mean μ and the standard deviation σ of the elevation values along the elevation gap borders are computed. If σ is low, the gap is merged with the adjacent 3-D region with the closest elevation to μ. If σ is high, the gap may correspond to a hidden part related to a discontinuity, and therefore the gap is kept. This process allows us to locally interpolate undefined points where there is no ambiguity.

An adjacency graph $\mathcal{G}$ is then created with the altimetric regions as nodes, linked by adjacency relations between regions.

As the DEM is not completely dense, adjacency relations are not complete. Some of them are missing because of the remaining elevation gaps due to occlusion areas. In order to remedy this problem, the graph is completed by involving the geometry of the views into the neighborhood definition. More precisely, the usual neighborhood relations based on adjacency are extended in the following way: two altimetric regions are neighbors in the graph not only when they are adjacent in the object space, but also when their projections onto the left or the right image plane are adjacent (see the example of Fig. 7). This extension creates neighborhood relations over elevation gaps due to occlusions. Through this process, the structure of the graph is not affected by hidden areas without 3-D information.

3.2. Binary Classification of Nodes as "Ground" or "Above-Ground"

We next label each node of the adjacency graph $\mathcal{G}$ as ground or above-ground. This label is called the *nature* of the node. According to the definition of an above-ground region given in Section 2.3, the nature of a node depends on the node itself and on its neighbors. Thus the decision cannot be taken at a local level only, but requires information taken over extended regions. This is important in order to detect everything above ground, for instance low above-ground objects located on a sloping ground

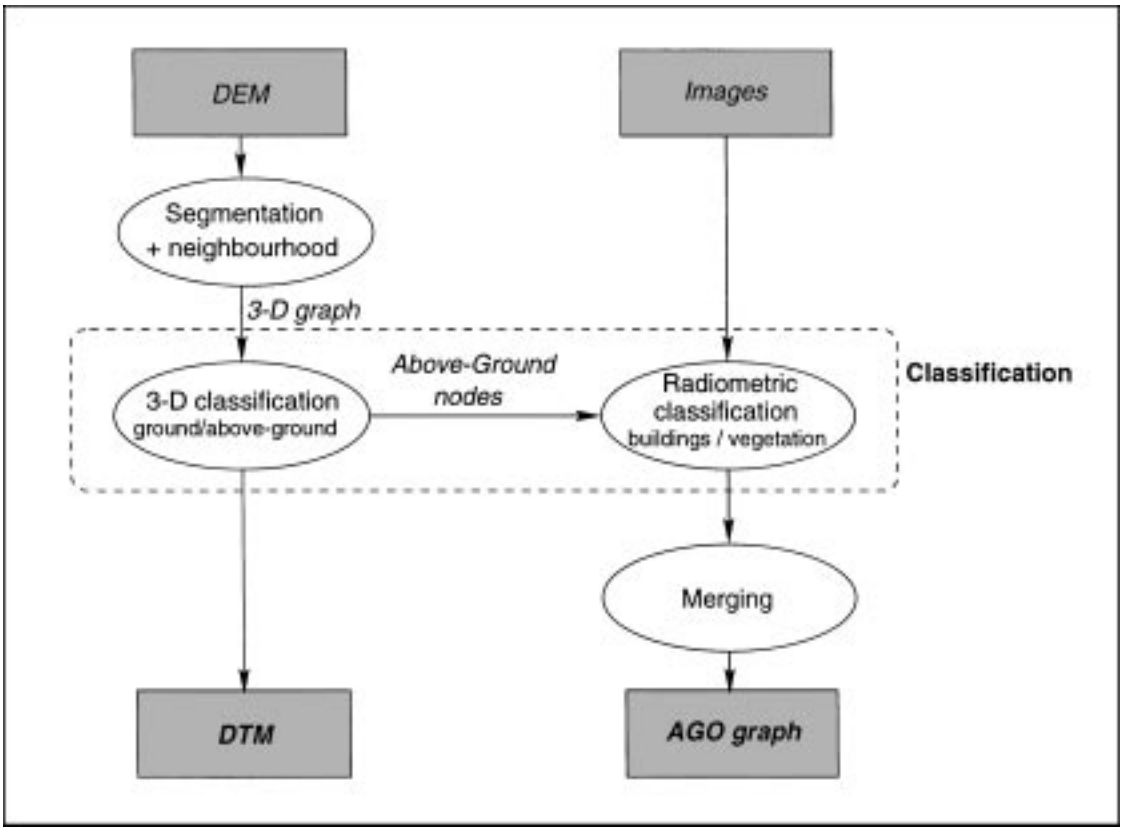

FIG. 6. Block diagram of the classification process for the detection of above-ground objects. The analysis of both the DEM and the radiometry leads to the computation of a DTM and an AGO graph.

surface or surrounded by higher regions (like the examples of Fig. 3). Markov random field models represent an adequate way to encode and to manipulate spatial interations between nodes, and they provide an elegant solution to optimization problems depending on local interactions [36].

3.2.1. Markovian model. The Markovian hypothesis assumes that a local conditioning is sufficient to determine the distribution of a random variable on a predefined set, whose elements are called *sites*. Let us consider that an input image is the realization of a random field X, called the *observation field*. One wants to extract from the observation a new image as the realization of a random field Y called the *descriptor field*. The descriptor field Y is said to be *Markovian* if the value taken in a site s only depends on the configuration of its neighboring sites (observation and descriptor values). The optimization problem is then be formulated as a minimization of a potential.

In our case, sites and neighborhood relations are given by the nodes and arcs of the 3-D graph $\mathcal{G}$ (irregular meshes). The observation field is defined by node elevations and it is denoted by H. The descriptor field is binary (ground, above-ground) and denoted by N. N is assumed to be Markovian conditioned on H, meaning that given H, the value of N in a site s only depends on the neighboring sites of s. This hypothesis is consistent with the definition of above-ground given previously.

Thus, the estimation of N can be formulated as the minimization of a potential funtion $U_{\mathcal{G}}(N \mid H)$ given by

$$U_{\mathcal{G}} = \alpha_d U_d + \alpha_c U_c. \tag{1}$$

- U_d is the *data attachment potential*, linking observation H and descriptor N; it is defined in each site s by a function $V_d(s)$ described in Section 3.2.3;
- U_c is the *contextual potential* describing the consistency of the descriptor H in view of the considered neighborhood; it is defined for each pair of neighboring sites (s, s') by a function $V_c(s, s')$ described in Section 3.2.2 (the neighborhood relations have been limited to the 2-order cliques).
- α_d and α_c are weighting parameters on U_d and U_c such as $\alpha_d + \alpha_c = 1$.

Let us note that $h(s)$ and $n(s)$ are the values of H and N, respectively, taken in a site s. There are two possible values for $n(s)$: $n(s) = Gnd$ (ground site) and $n(s) = Abv$ (above-ground site).

3.2.2. Contextual potential $V_c(s, s')$. It is assumed that $V_c(s, s')$ only depends on $n(s), n(s')$ and on the difference in height:

$$\delta(s, s') = h(s) - h(s'). \tag{2}$$

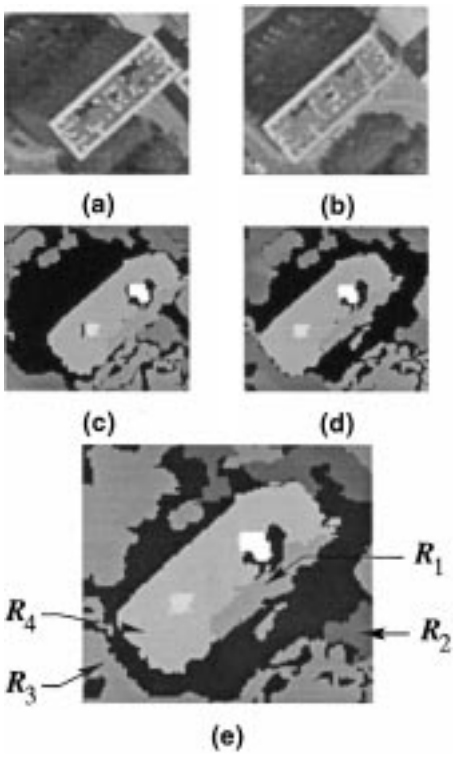

FIG. 7. Neighborhood extension over occlusions. (a) and (b) Left and right images of a high building and its close neighbourhood; (c) and (d) corresponding altimetric regions, projected onto each image referential; (e) corresponding altimetric regions in the object (orthographic) referential. Although the regions R_1 and R_2 (respectively, R_3 and R_4), are not neighbors in the object referential, neighborhood links are created between them, because of the adjacency of the corresponding regions projected onto the left (resp., the right) referential.

Let us note $V_c(s, s') = V_c^{n,n'}(\delta)$. There are four cases to consider according to the value of the pair (n, n'). The appearance of these functions is shown in Fig. 8 and exact definitions can be found in [35].

The functions are parameterized by a critical value δ_0, which is mentionned in the definition of above-ground: it is the maximal difference in height between two neighboring ground nodes. Therefore,

$$\delta \leq \delta_0 \Leftrightarrow V_c^{Gnd,Gnd}(\delta) \leq V_c^{Abg,Gnd}(\delta). \tag{3}$$

The parameter δ_0 characterizes the above-ground: the smaller it is, the more numerous the detected above-ground is. Its minimal value is related to the growing threshold δ_g used during the segmentation step by the following relation: $\delta_0 \geq 2\delta_g$.

The potential is symmetrical with respect to s and s':

$$V_c^{n,n'}(\delta) = V_c^{n',n}(-\delta) \quad \forall\delta\, \forall(n, n'). \tag{4}$$

Moreover, $V_c^{Abg,\,Abg}(\delta) = -1\, \forall\delta$ because urban scenes can often be found with two adjacent above-ground objects having very different heights.

The functions are made of arcs of a Gaussian to guarantee continuity and derivability. Moreover, the derivative is zero around the critical value δ_0 (or $-\delta_0$) for the stability of the classification process.

3.2.3. Data attachment potential $V_d(s)$. There are two cases to consider according to the nature of s. If s is labeled as ground, then $V_d(s)$ is an increasing function of the elevation $h(s)$. On the contrary, if s is labeled as above-ground, then $V_d(s)$ is a decreasing function of $h(s)$. The corresponding functions are shown in Fig. 9.

The functions are parameterized by the elevation values $h_0(s)$, $h_1(s)$, and $h_2(s)$. The elevation $h_0(s)$ is a critical value above which $V_d(s)$ is favorable to the above-ground nature. Parameters $h1(s)$ and $h2(s)$ determine the width of both arcs of a Gaussian. Those three parameters are defined according to the local height and slope of the ground:

$$\begin{cases} h_0(s) = h_{G_{avg}}(s) + \delta_0 \\ h_1(s) = \min(h_{G_{min}}(s) + \delta_0, h_{G_{avg}}(s)) \\ h_2(s) = \max(h_{G_{max}}(s) + \delta_0, h_{G_{avg}}(s) + 2\delta_0), \end{cases} \tag{5}$$

where $h_{G_{avg}}(s)$, $h_{G_{min}}(s)$, and $h_{G_{max}}(s)$ are, respectively, the average, minimal, and maximal elevations of the ground surface at the site s.

An estimation of the topographic surface h_G is thus necessary in order to provide the parameter values $h_{G_{avg}}(s)$, $h_{G_{min}}(s)$, and $h_{G_{max}}(s)$ in each site s. Given a set of 3-D points assumed to belong to the ground, we suggest performing this estimation by an appropriate sampling followed by an interpolation step.

More precisely, the elevation image is subdivided into windows of similar size. Then one point per window is selected: it is a ground point located at the most frequent elevation of the window and as close as possible to the center. The set of the selected points over the elevation image is processed by a Delaunay triangulation (in 2-D). Elevations are finally interpolated assuming that each triangle describes a planar surface [37].

An example of a triangulation and the corresponding DTM are shown in Fig. 10.

3.2.4. Optimization process. Above-ground detection and DTM estimation interact according to the scheme of Fig. 11. On one hand, the ground elevations are necessary to parameterize the data attachment potential function. On the other hand, the above-ground localization is important to filter elevation data before the sampling for DTM computation.

Therefore, we suggest the iterative optimization process shown in Fig. 12. The DTM is computed and updated according to the current classification of nodes, and the global energy is minimized over the graph using the DTM as *a priori* information. The nodes are initialized as ground, and the process is iterated until stability.

The convergence of the process is not guaranteed, but experiments have shown that a stable state was reached most of the

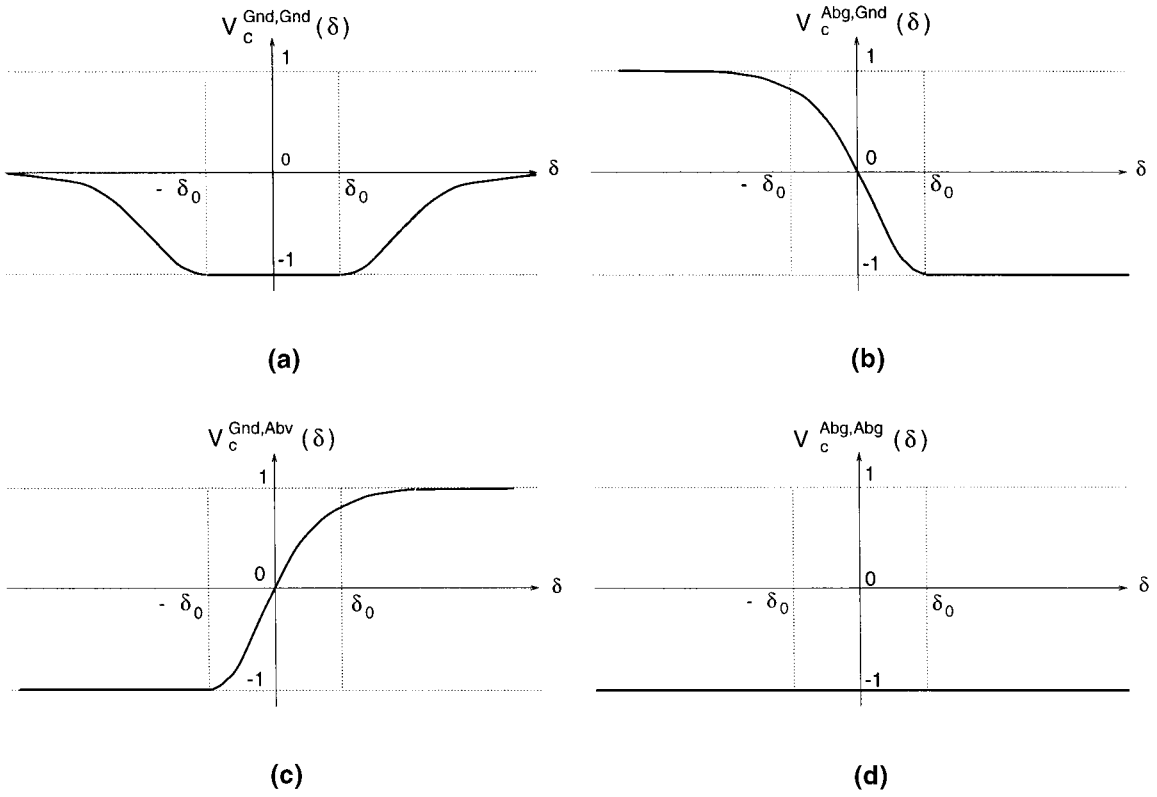

FIG. 8. Contextual potential functions $V_c^{n,n'}(\delta)$ for each value of (n, n') versus the difference in height between two sites $\delta(s, s') = h(s) - h(s')$. (a) Both sites s and s' are supposed to be on the ground; (b) The site s belongs to the above-ground, the site s' is on the ground; (c) The site s is on the ground, the site s' belongs to the above-ground; (d) Both sites s and s' belong to the above-ground.

time. Still an oscillation between two or three different states can occur. In this case, the nodes whose classification is not stable (always less than 0.5% of the nodes in our experiments) are always small in size and they correspond to really ambiguous regions. In our experiments they are detected and labeled as "ground."

3.3. Characterization of Above-Ground Nodes as Building or Vegetation

When each node of the 3-D graph has been classified as ground or above-ground, the radiometric information is used to separate above-ground nodes into two classes: building and vegetation.

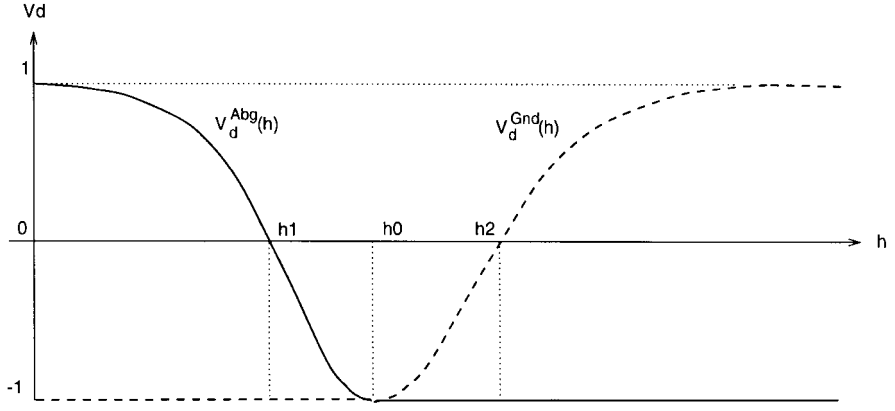

FIG. 9. Data attachment potential fuction $V_d(s)$ in a site s versus its height $h(s)$. If $h(s) > h_0(s)$ (resp., $h(s) < h_0(s)$), the site s is more likely to be an above-ground (resp., ground) node.

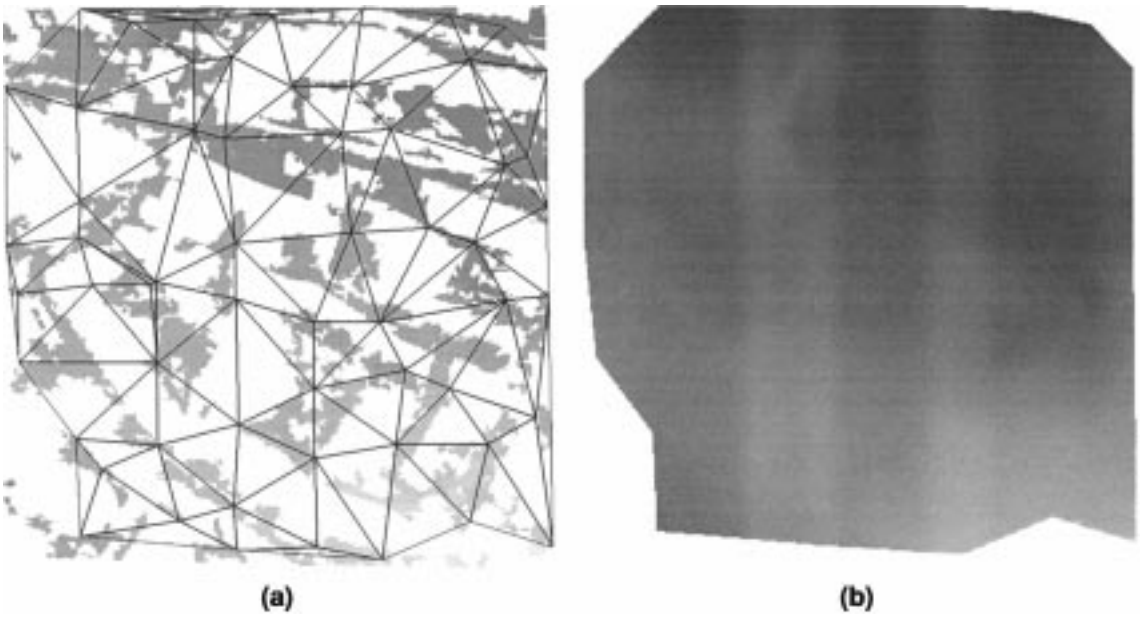

FIG. 10. DTM estimation: (a) Delaunay triangles produced from the set of selected 3-D ground points; (b) DTM after interpolation of the ground elevations.

For that purpose, an analysis of the left radiometric image is first performed, driven by the above-ground location. We use a textural measure described in [38]. It is a local measure based on the entropy of the radiometric gradient directions. For each pixel of the gray-level image, the histogram of the gradient directions (modulo $\frac{\pi}{2}$) is computed over a centered neighborhood. The entropy of the probability density derived from the histogram is then assigned to the central pixel. A low entropy value denotes a main direction of the gradient near the pixel, whereas a high entropy value means that the neighborhood is not structured. The threshold on the entropy values is computed automatically through an appropriate resampling described in [38].

The textural measure is associated to the altimetric classification and to additional radiometric and topological criteria, in order to segment the gray-level image into three classes: ground, building, and above-ground vegetation. The additional radiometric and topological criteria express general knowledge about the scene such as the facts that vegetation is dark or that small above-ground vegetation areas surrounded by lower buildings are unlikely (see [35] for details). These criteria are important to distinguish vegetation from roof superstructures for instance.

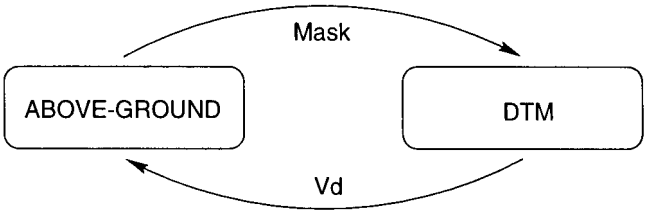

FIG. 11. Interaction between above-ground detection and DTM estimation.

Finally, each above-ground node of the graph is associated to the most frequent label of its corresponding pixels (after projection onto the gray-level image).

3.4. Merging into Above-Ground Objects

As a final step to the segmentation process, AGOs are produced by merging together neighboring nodes belonging to the same class and separated by a small difference in height. The topology between nodes is propagated to the object level.

AGOs are then the nodes of a new graph called an *AGO graph* (see Fig. 13), which captures not only altimetric and thematic information, but also topological relations between the objects.

4. EXPERIMENTATION AND DISCUSSION

The AGOs and the classification obtained from the data of Fig. 1 are shown in Figs. 15 and 17a. Figure 14 shows another example of aerial image and corresponding DEM, and Figs. 16 and 17b show the results of the classification process.

Even with a ground slope of locally 15%, about 95% of the above-ground surface have been detected and correctly classified: extended above-ground, dense urban aggregates, buildings with large variation in height, as well as small objects. The final segmentation, based on altimetric, topological, and radiometric criteria, has provided relevant areas of interest. The borders are not very accurate and sometimes locally irregular because of the initial DEM, but the global shape of objects has been preserved. Besides the classification itself, the process provides a DTM of the scene and a symbolic 3-D representation of the scene as an AGO graph, which can be used to exploit interactions between neighboring objects.

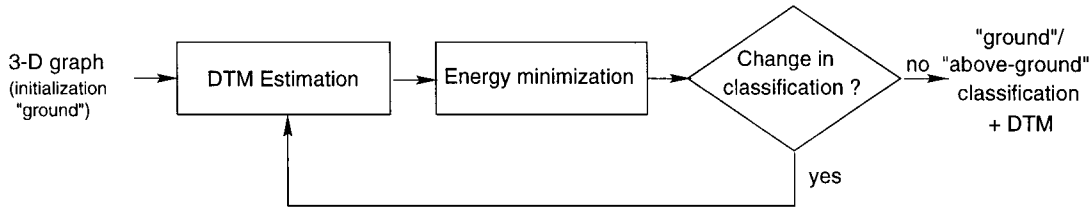

FIG. 12. Iterative process for the binary classification ground/above-ground.

These results show the relevancy of a simple and generic model for AGO. The detection requires no geometric knowledge and few parameters. The processing time needed for a 1000×1000 size image is about 10 min on a station alpha AXP 150 MHz. It is almost equally distributed between the graph creation (extension of neighborhood relations), the binary classification of the nodes as ground or above-ground, and the textural measure based on entropy values.

A quantitative assessment on five different stereo pairs (various scenes and various resolution photographs) covering a total area of about 1 km^2 has been performed with a unique set of parameter values ($\delta_g = 1$ m, $\delta_0 = 2$ m, $\alpha_c = \alpha_c = 0.5$).

It has revealed that between 91.5 and 97% of the above-ground surface was correctly detected (false detection between 2 and 11%). Only small objects like isolated trees or very small houses are sometimes lost. False detections of the above-ground surface often come from matching errors due to homogeneous radiometry. The quality of the detection is stable over the experimented range of image resolution (between 20 cm and 1 m per pixel).

The study has also shown that between 87.6 and 94% of the buildings were correctly detected (false detection 13%). These results deteriorate a little when the image resolution gets poorer than 50 cm per pixel, as the linear characteristics of buildings become less clear. Confusion between buildings and trees occurs when the main direction of a roof does not appear clearly in the image (cases of a few small oblic roofs) or, more frequently, when an inappropriate neighborhood is taken into account in the local textural measure computation. This could be avoided by using both images of the stereo pair, since the spatial context of objects can vary a lot between the two points of view, especially in the neighborhood of AGOs.

5. CONCLUSION

A new approach for the 3-D reconstruction of complex urban scenes from a pair of mid-resolution aerial images has been proposed. It consists of detecting and classifying the above-ground objects of the scene, buildings, and vegetation. In order to cope with the complexity and the diversity of urban environments, only low-level 2-D and 3-D properties have been used, through the computation and the segmentation of a DEM and a local analysis of radiometry. The method provides a symbolic representation of the scene combining different kinds of

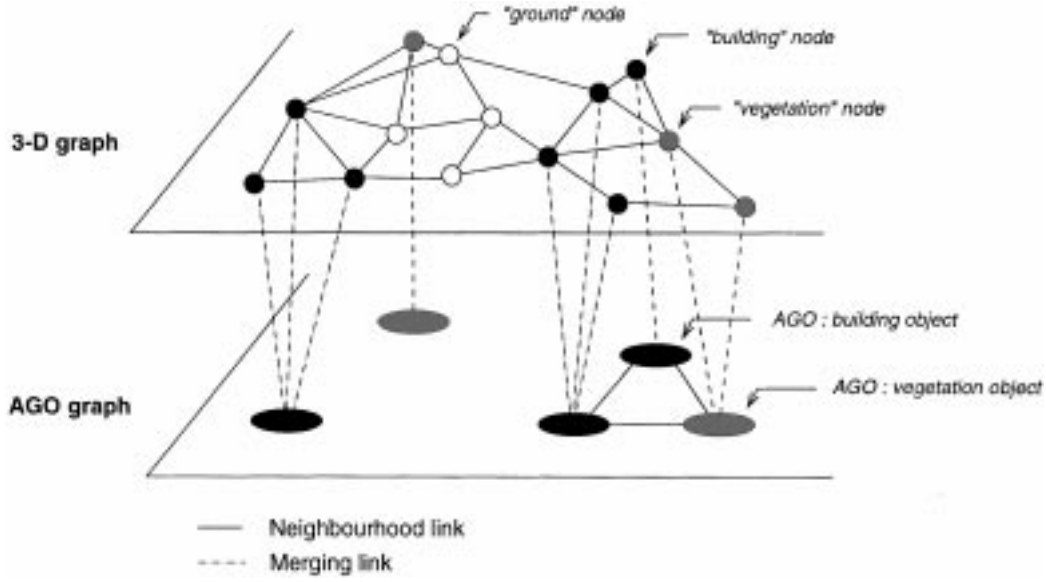

FIG. 13. Creation of the AGO graph by merging neighboring "above-ground" nodes.

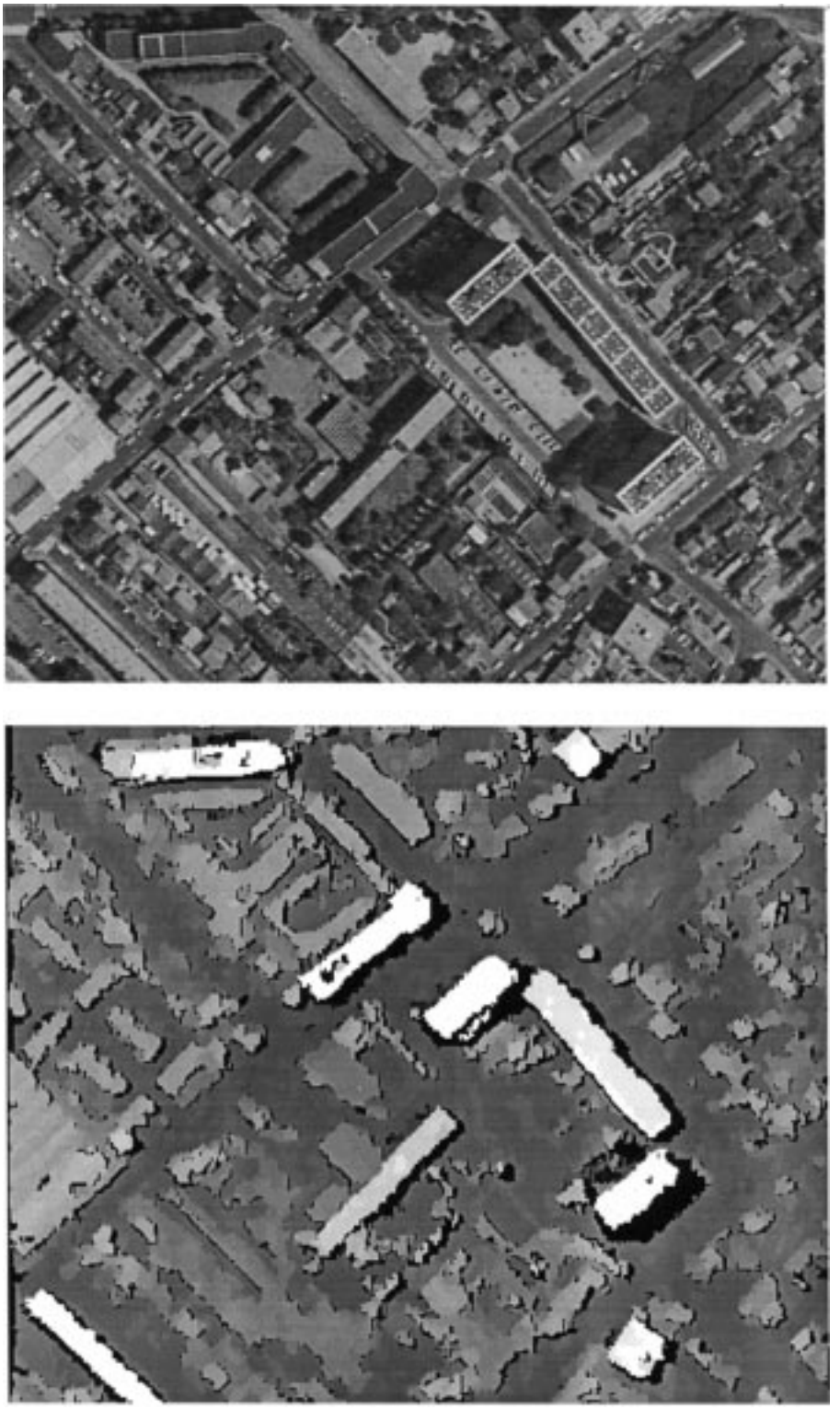

FIG. 14. Aerial image (size 860 × 700 pixels, same characteristics as the images of Fig. 1) and the corresponding DEM. The DEM has been automatically computed from two images (only one of them is shown here).

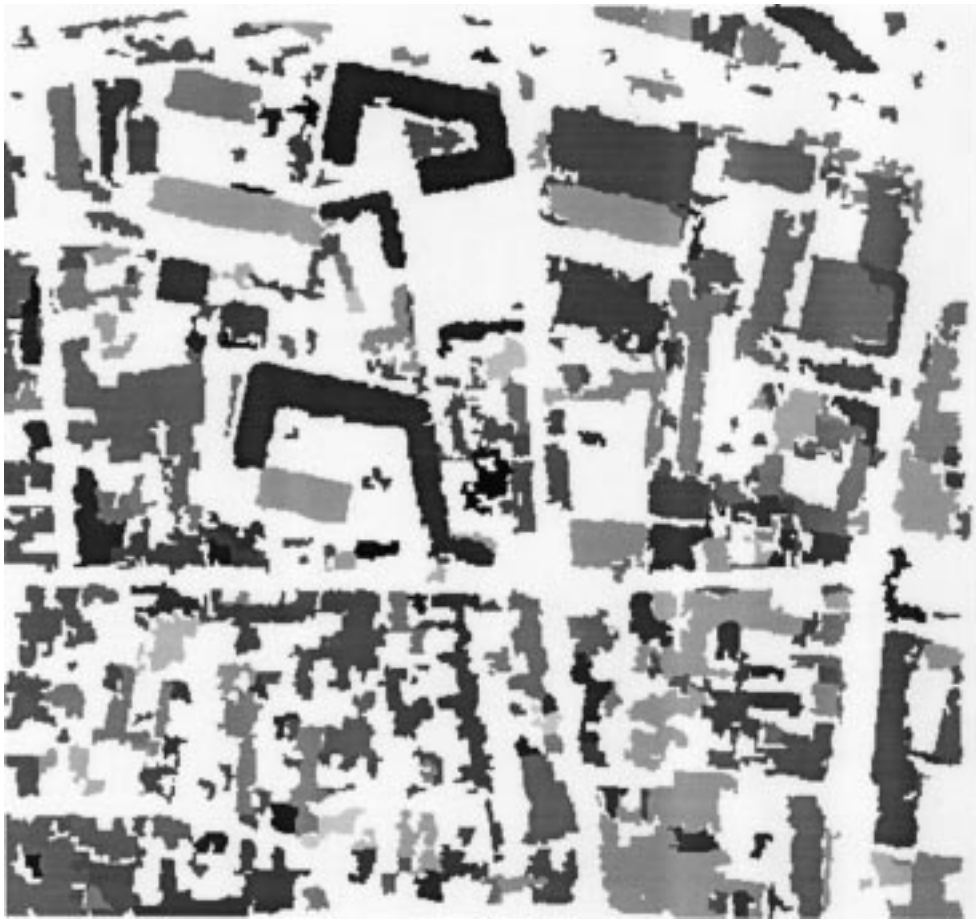

FIG. 15. Result of the segmentation computed from the input images of Fig. 1 and the corresponding DEM of Fig. 5 ($\delta_g = 1$ m, $\delta_0 = 2$ m, $\alpha_c = \alpha_c = 0.5$). Each above-ground object is represented by a random gray level.

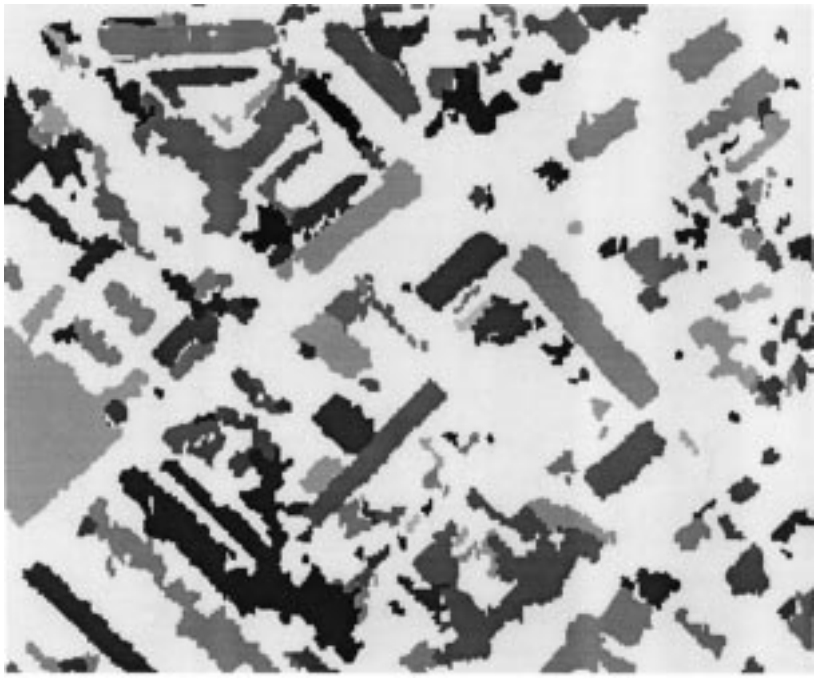

FIG. 16. Result of the segmentation computed from the input data of Fig. 14 ($\delta_g = 1$ m, $\delta_0 = 2$ m, $\alpha_c = \alpha_c = 0.5$). Each above-ground object is represented by a random gray level.

(a)

(b)

FIG. 17. Result of the classification: red hatched areas are classified as building and green hatched areas as vegetation (left areas are classified as ground). (a) results for the input data of Fig. 1; (b) results for the input data of Fig. 14.

information: thematic, geometric, radiometric, and contextual. It also provides a DTM of the ground surface.

Results show that the method is reliable and robust to scene and image variability. The novelty of the approach lies with its ability to cope with very different contexts, which is of prime importance when one deals with urban environments. The robustness has been achieved by involving a generic 3-D model and complementary low-level information. The use of a Markovian model within an iterative optimization process enables us to simultaneously take the elevation of the ground surface and local differences in height into account. Therefore, the method does not require a fine tuning of parameter values.

The information should enable a fine 3-D reconstruction of the scene to be realized, locally adapted to the objects and to their context. The selection of spatial features related to areas of interest will reduce the combinatorial complexity and hence the risk of errors. The local information which has been accumulated during the process can be used in many ways: hypothesis generation, selection of appropriate models, relevant initialization, control on parameters, etc. We therefore believe that the focusing strategy proposed in this paper will prove extremely useful in processing complex urban scenes.

ACKNOWLEDGMENTS

We are very grateful to Olivier Dissard, from the IGN, for his assistance and many helpful discussions. Financial support for this work was provided by the IGN.

REFERENCES

1. Y.-T. Liow and T. Pavlidis, Use of shadows for extracting buildings in aerial images, *Comput. Vision Graphics Image Process.* **49**, 1990, 242–277.
2. J. A. Shufelt and D. M. McKeown, Fusion of monocular cues to detect man-made structures in aerial imagery, *CVGIP: Image Understanding* **57**, 1993, 307–330.
3. T. Kim and J. P. Muller, Automated building height estimation and object extraction from multi-resolution imagery, in *Integrating Photogrammetric Techniques with Scene Analysis and Machine Vision II, Orlando, FL, 1995*, SPIE Vol. 2486, pp. 267–276.
4. C. Lin and R. Nevatia, Building detection and description from a single intensity image, *Comput. Vision Image Understanding* **72**, 1998, 101–121.
5. A. Huertas and R. Nevatia, Detecting buildings in aerial images, *Comput. Vision Graphics Image Process.* **41**, 1988, 131–152.
6. R. B. Irvin and D. M. McKeown, Methods for exploiting the relationship between buildings and their shadows in aerial imagery, *IEEE Trans. Systems Man Cybernet.* **19**, 1989, 1564–1575.
7. J. C. McGlone and J. A. Shufelt, Projective and object space geometry for monocular building extraction, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Seattle, 1994*, pp. 54–61.
8. R. Mohan and R. Nevatia, Using perceptual organization to extract 3-d structures, *IEEE Trans. PAMI* **11**, 1989, 1121–1139.
9. H. J. Lee and W. L. Lei, Region matching and depth finding for 3-d objects in stereo aerial photographs, *Pattern Recognition* **23**, 1990, 81–94.
10. U. Stilla and K. Jurkiewicz, Structural 3D-analysis of urban scenes from aerial images, in *XIX Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Vienna, July 1996*, Vol. 31, pp. 832–838.
11. 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 *XVIII Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Munich, September 1994*, Vol. 30, pp. 165–172.
12. T. Quiguer, Rectangular building 3d reconstruction in urban zones, in *XIX Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Vienna, July 1996*, Vol. 31, pp. 657–662.
13. M. Roux and D. M. McKeown, Feature matching for building extraction from multiple view, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Seattle June 1994*.
14. U. Weidner and W. Förstner, Towards automatic building extraction from high-resolution digital elevation models, *ISPRS J. Photogram. Remote Sensing* **50**, 1995, 38–49.
15. S. Noronha and R. Nevatia, Detection and description of buildings from multiple aerial images, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Puerto Rico*, 1997, pp. 588–594.
16. N. Paparoditis, M. Cord, M. Jordan, and J.-P. Cocquerez, Building detection and reconstruction from mid- and high-resolution aerial imagery, *Comput. Vision Image Understanding* **72**, 1998, 122–142.
17. M. Berthod, L. Gabet, G. Giraudon, and J. L. Lotti, High-resolution stereo for the detection of buildings, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images, Ascona, Switzerland* (A. Grün, O. Kübler, and P. Agouris, Eds.), pp. 135–144, Birkhauser, Basel, 1995.
18. N. Haala and M. Hahn, Data fusion for the detection and reconstruction of buildings, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images, Ascona, Switzerland* (A. Grün, O. Kübler, and P. Agouris, Eds.), pp. 211–220, Birkhauser, Basel, 1995.
19. H. Wiman and P. Axelsson, Finding 3D-structures in multiple aerial images using lines and regions, in *XIX Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Vienna, July 1996*, Vol. 31, pp. 953–959.
20. Y. Lechervy, C. Louis, and O. Monga, Crestline contribution to the automatic building extraction, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images (II), Ascona, Switzerland* (A. Grün, E. P. Baltsavias, and O. Henricsson, Eds.), pp. 161–172, Birkhäuser, Basel, 1997.
21. U. Weidner, An approach to building extraction from digital surface models, in *XIX Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Vienna, July 1996*, Vol. 31, pp. 924–929.
22. O. Henricsson, F. Bignone, W. Willuhn, F. Ade, O. Kübler, E. Baltsavias, and S. Mason, Project AMOBE: Strategies, current status and future work, in *XIX Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Vienna, 1996*, Vol. 31, pp. 321–330.
23. T. Moons, D. Frère, J. Vandekerckhove, and L. Van Gool, Automatic modelling and 3d reconstruction of urban house roofs from high resolution aerial imagery, in *Proc. of 5th European Conf. on Computer Vision, Freiburg, Germany, 1998*, pp. 410–425.
24. O. Henricsson, The role of color attributes and similarity grouping in 3-d building reconstruction, *Comput. Vision Image Understanding* **72**, 1998, 163–184.
25. S. Girard, P. Guérin, H. Maître, and M. Roux, Building detection from high resolution colour images, in *SPIE Europto Image and Signal Processing for Remote Sensing IV, Barcelona, 1998* (S. Serpico, Ed.), Vol. 3500, pp. 278–289.
26. C. Baillard and A. Zisserman, Automatic reconstruction of piecewise planar models from multiple views, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Fort Collins, June 1999*.
27. M. Fradkin, M. Roux, H. Maître, and U. M. Leloglu, Surface reconstruction from multiple aerial images in dense urban areas, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recongnition, Fort Collins*, 1999.

28. M. Roux, Cooperative analysis of maps and aerial images for urban scene description, in *SPIE Europto: Image and Signal Processing for Remote Sensing III, London, September 1997*, Vol. 3217, pp. 254–267.
29. M. Pasko and M. Gruber, Fusion of 2D GIS data and aerial images for 3D building reconstruction, in *XIX Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Vienna, 1996*, Vol. 31, pp. 257–260.
30. C. Brenner and N. Haala, Fast production of virtual reality city models, in *ISPRS Comm. IV Symposium, "GIS between Visions and Applications," Stuttgart, 1998, IAPRS*, Vol. 32, part 4, pp. 77–84.
31. E. Baltsavias, S. Mason, and D. Stallman, Use of DTMs/DSMs and orthoimages to support building extraction, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images, Ascona, Switzerland* (A. Grün, O. Kübler, and P. Agouris, Eds.), pp. 199–210, Birkhauser, Basel, 1995.
32. N. Haala, Detection of buildings by fusion of range and image data, in *XVIII Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Munich, September 1994*, Vol. 30, pp. 341–346.
33. W. Eckstein and C. Steger, Fusion of digital terrain models and texture for object extraction, in *2nd Int. Airbone Remote Sensing Conf. and Exhibition, 1996*, pp. 1–10.
34. C. Hug, Extracting artificial surface objects from airbone laser scanner data, in *Automatic Extraction of Man-Made Objects from Aerial and Space Images (II), Ascona, Switzerland* (A. Grün, E. P. Baltsavias, and O. Henricsson, Eds.), pp. 203–212, Birkhäuser, Basel, 1997.
35. C. Baillard, *Analyse d'images aériennes stéréoscopiques pour la restitution 3-D des milieux urbains*, Ph.D. thesis, ENST 97-E-018, Paris, October 1997. [Available at http://www-isis.enst.fr/Kiosque/theses/MANUSCRITS/]

35a. C. Baillard and O. Dissard, A stereo matching algorithm for urban digital elevation models, *Photogrammetric Engineering and Remote Science*, to appear.

36. R. Azencott, Image analysis and Markov fields, in *Int. Conf. on Ind. and Appl. Math. SIAM, Paris, 1988*.
37. M. Roux, J. Lopez-Krahe, and H. Maître, Automatic digital terrain model generation using aerial images and maps, in *XIX Congress of ISPRS, Comm. III, Int. Archives of Photogrammetry and Remote Sensing, Vienna, July 1996*, Vol. 31, pp. 697–702.
38. C. Baillard, O. Dissard, O. Jamet, and H. Maître, Extraction and characterization of above-ground areas in a peri-urban context, in *Mapping Buildings, Roads and Other Man-Made Structures from Images, Proc. of the IAPR TC7 Workshop on Remote Sensing and Mapping, Graz, Austria, September 1996*, pp. 159–174.