Large-Scale Building Reconstruction Through Information Fusion and 3-D Priors

Konstantinos Karantzalos, *Member, IEEE*, and Nikos Paragios, *Senior Member, IEEE*

***Abstract*—In this paper, a novel variational framework is introduced toward automatic 3-D building reconstruction from remote-sensing data. We consider a subset of building models that involve the footprint, their elevation, and the roof type. These models, under a certain hierarchical representation, describe the space of solutions and, under a fruitful synergy with an inferential procedure, recover the observed scene's geometry. Such an integrated approach is defined in a variational context, solves segmentation both in optical images and digital elevation maps, and allows multiple competing priors to determine their pose and 3-D geometry from the observed data. The very promising experimental results and the performed quantitative evaluation demonstrate the potentials of our approach.**

***Index Terms*—Level sets, modeling, object detection, recognition, registration, segmentation, variational methods.**

I. Introduction

THREE-DIMENSIONAL building and landscape models are of great interest for various engineering applications such as urban and rural planning, updating geographic information system (GIS), augmented reality, 3-D visualization, virtual tourism, location-based services, navigation, wireless telecommunications, disaster management, noise, and heat and exhaust-spreading simulations. All are actively discussed in the computer vision and geoscience scientific community, and some of them have already entered or are expected to enter the market soon. The prohibitively high costs of generating manually such models explain the emergency toward automatic approaches. Furthermore, it should be noted that the required output spatial accuracy is of major importance, particularly at large scales, as it designates the method's operational functionality, performance, and success.

In order to obtain such a 3-D vector description of a scene's geometry, apart from single 2-D aerial and satellite images, 3-D information—digital elevation models (DEMs)–is also required. DEMs can be obtained indirectly with classical photogrammetric multiview-stereo techniques [1] or more recently, using emerging airborne or spaceborne active sensors like Light Detection and Ranging (LIDAR), also known as airborne laser scanning, and interferometric synthetic aperture radar (INSAR). The collection of DEMs from such active sensors is increasing rapidly as these technologies become more widely available and cost effective, contrary to the indirect image-based multiview-stereo reconstruction methods [2]. In particular, image matching and accurate breakline positioning are not trivial tasks and become cumbersome as spatial resolution gets higher over complex scenes or urban regions and in untextured areas or at depth discontinuities [3]. Both LIDAR and INSAR principles allow building-detection applications with an advantage on LIDAR in terms of spatial resolution and potential nadir-view acquisitions, and with an advantage on INSAR in terms of the technology's robustness in weather conditions [4].

Despite the recent intensive research toward 3-D building extraction and reconstruction based on various remote-sensing data and several model-free or model-based procedures ([5]–[13] and the references therein), we are still far from the goal of the initially envisioned fully automatic and accurate reconstruction systems [14], [15]. Processing remote-sensing data still poses several challenges.

On the one hand, intensity images, 2-D projections of the real 3-D world, are inherently ambiguous, and shadows or occlusions frequently occur. Edge-, line-, corner-, and junction-detection techniques as well as purely image-driven (edge or region-based) segmentation techniques usually fail to operate effectively due to the misleading low-level information. Algorithms (like [16]–[20]) that were designed to extract and reconstruct buildings based only on purely image-driven functions and step-by-step procedures possess native limitations. For several scenes, moreover, DEMs from several sources are often available. Thus, sophisticated approaches should be able to adapt to every given situation and be able to process images, point clouds, or height/depth/disparity maps.

On the other hand, processing LIDAR or INSAR data involves, in general, oblique view acquisitions, limited resolution of the samples near surface edges, presence of noise due to errors from the GPS/inertial navigation system, other registration errors, poor reflectivity properties of some surfaces, shadowing/layover effects, multipath backscattered signals, and speckle [4]. Although, they are nowadays holding a significant position in the market that is increasing rapidly, in most cases, the overall spatial resolution of the cost-effective acquisitions is usually lower than the nowadays commercially available high-resolution aerial and satellite imagery [14, Fig. 1(a) and (b)].

Manuscript received February 16, 2009; revised July 16, 2009, September 16, 2009, and October 31, 2009. First published March 18, 2010; current version published April 21, 2010. This work was supported in part by the Conseil General de Hauts-de-Seine and in part by the Region Ile-de-France under the TERRA NUMERICA Grant of the Pole de competitivite CapDigital.

K. Karantzalos is with the Remote Sensing Laboratory, National Technical University of Athens, Zografou Campus, 15780 Athens, Greece (e-mail: karank@central.ntua.gr).

N. Paragios is with the Medical Imaging and Computer Vision Group, Applied Mathematics and Systems Laboratory, Department of Applied Mathematics, Ecole Centrale de Paris, 92295 Chatenay-Malabry, France, and also with the Institut National de Recherche en Informatique et en Automatique (INRIA) Saclay Ile-de-France, 91893 Orsay, France (e-mail: nikos.paragios@ecp.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2009.2039220

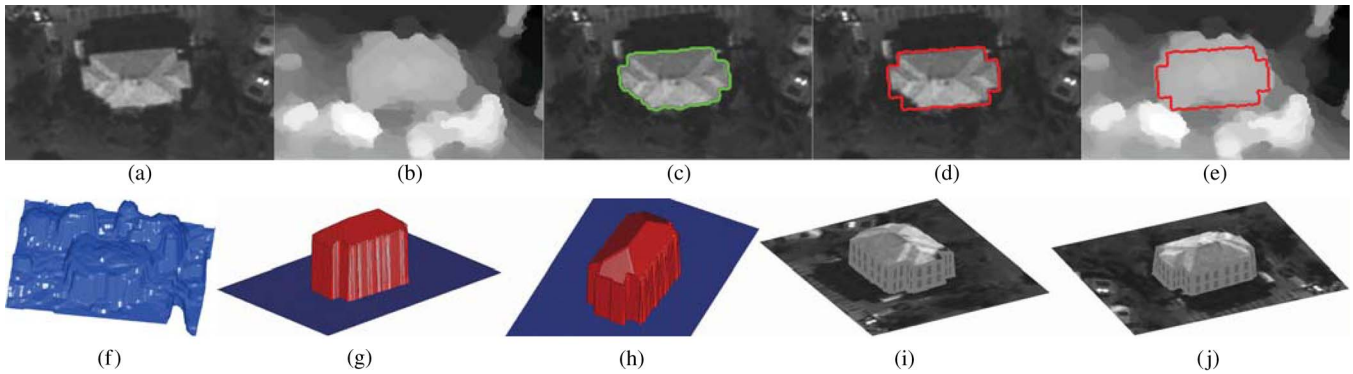

Fig. 1. Proposed variational framework is constrained from the cue with the higher spatial resolution and can accurately extract and reconstruct scene buildings overcoming shadows, occlusions, and low-level misleading information. (First row) (a) Original image. (b) DEM. (c) Detected building footprints from a conventional purely data-driven segmentation. Detected building footprints from the developed algorithm superimposed (d) to the initial image and (e) to the DEM. (Second row) (h) DEM's 3-D visualization and (i)–(k) different 3-D views after the proposed reconstruction process.

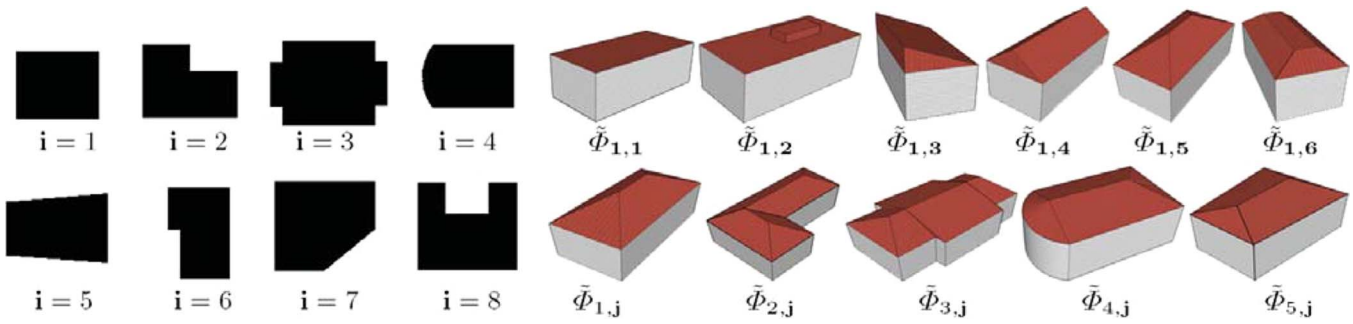

Fig. 2. Three-dimensional prior building models ($\tilde{\Phi}_{\mathbf{i},\mathbf{j}}$): $\mathbf{i}$ determines the shape of the building's footprint and $\mathbf{j}$ its roof type. (Left) Eight different types of building footprints. (Right top) The family $\tilde{\Phi}_{1,\mathbf{j}}$ of building priors that have a rectangular footprint ($\mathbf{i} = 1$). (Right bottom) The family $\tilde{\Phi}_{\mathbf{i}=\mathbf{1:5},\mathbf{j}}$ of prior models.

Thus, algorithms (like [2], and [21]–[23]) that were designed to reconstruct buildings exclusively from DEMs are—mostly or partly—based on corner-, junction-, edge- or line-detection processes, and step-by-step procedures possess native limitations particularly in terms of spatial accuracy.

In this paper, we aim to address the aforementioned challenges by introducing a novel variational framework toward large-scale building reconstruction through information fusion and grammar-based building priors. Multiple 3-D competing priors are considered transforming reconstruction to a labeling and an estimation problem. In such a context, we fuse images and DEMs toward recovering a 3-D model. Our formulation allows data with the higher spatial resolution to constrain properly the footprint detection in order to achieve the optimal spatial accuracy (Fig. 1, top). Therefore, we are proposing a variational function that encodes a fruitful synergy between observations and multiple 3-D grammar-based building models. Our models refer to a grammar, which consists of typologies of 3-D shape priors (Fig. 2). In such a context, first, one has to select the most appropriate model and then determine the optimal set of parameters aiming to recover the scene's geometry (Fig. 1, bottom). The proposed objective function consists of two segmentation terms that guide the selection of the most appropriate typology and a third DEM-driven term which is being conditioned on the typology. Such a prior-based recognition process can segment both rural and urban regions (similar to [2]) but is able as well to overcome detection errors caused by the misleading low-level information (like shadows or occlusions), which is a common scenario in remote-sensing data [Fig. 1(a) and (c)].

Our goal was to develop a single generic framework (with no step-by-step processes) that is able to efficiently account for multiple 3-D building extraction, no matter if their number or shape is *a priori* familiar or not. The motivation was to design an automated and generic solution based on the type of data that are nowadays most available and cost effective. In particular, we worked with aerial and satellite images of high and very high resolution (0.5–1.5 m ground resolution) and with elevation maps of medium and high resolution (1.0–3.0 m ground resolution). Doing multiview stereo, using simple geometric representations like 3-D lines and planes, merging data from ground sensors, or working with dense height data of very high ground resolution (< 0.6 m) was not our interest here. In addition, since usually for most sites, multiple aerial images are missing, our goal was to provide a solution even with the minimum available data, like a single panchromatic image and an elevation map, contrary to approaches that were designed to process multiple aerial images or multispectral information and cadastral maps (like in [17], [24], and [25]), which much eases the scene's classification. Moreover, contrary to [26], the here proposed variational framework does not require dense image-matching processes and *a priori* given 3-D line segments or a rough segmentation. The main contributions of this paper are the following.

1) We have developed a novel recognition-driven variational framework to address multiple 3-D building extraction

and reconstruction. It is an inferential approach that fuses optical images and digital elevation maps, is defined in a variational context, solves segmentation in both spaces, and allows multiple competing priors to determine their pose and 3-D geometry from the observed data.

2) We have introduced a grammar-based building representation to efficiently describe the space of solutions. By describing our numerous building models with a certain hierarchy and grammar and by formulating, respectively, our energy terms, the search space of solution during the optimization procedure has been narrowed effectively. Apart from new building models, other classes of terrain features can be added or removed from the database, controlling, respectively, the type of objects that can be addressed by the system.

The remainder of this paper is structured in the following way. In Section II, the introduced grammar-based representation of the prior building models is described. The proposed variational framework for multiple 3-D building extraction and reconstruction is detailed in Section III, along with a description of energy minimization and optimization steps. Experimental results and the performed quantitative evaluation are given in Section IV, and finally, conclusions and perspectives for future work are given in Section V.

II. Building Modeling Through a Hierarchical Grammar

Numerous 3-D model-based approaches have been proposed in literature. Statistical approaches [27], [28] aim to describe variations between the different prior models by measuring the distribution of the parameter space. These models are capable of modeling a building with rather repeating structure and of limited complexity. In order to overcome this limitation, methods using generic, parametric, polyhedral, and structural models have been considered [16]–[20], [22], [23]. The main strength of these models is their expressional power in terms of complex architectures. On the other hand, inference between the models and observations is rather challenging due to the important/high dimension of the search space. Consequently, these models can only be considered in a small number. More recently, procedural modeling of architectures has been introduced, as well as vision-based reconstruction, using mostly facade views [29]. Such a method recovers the 3-D geometry using an L-system grammar [30] which is a powerful and elegant tool for content creation. Despite the promising potentials of such an approach, one can claim that the inferential step that involves the derivation of model parameters automatically is still a challenging problem, particularly when the grammar is related to the building-detection procedure [31]–[34].

Hierarchical representations are a natural selection to address complexity while at the same time recover representations of acceptable resolution. Toward this end, a dictionary of basic shapes (intermediate and final ones) was employed, which uses a set of footprints that are parametric (the same footprint can produce numerous buildings with the same concept geometry); then, with an extrude rule, the volume of the building is generated (subject to certain parameters), another rule will split the main building from the roof (subject to certain parameters), and then it decomposes the roof into parts (subject to certain parameters). The employed vocabulary of basic shapes include the footprint, the volume, the main building, the roof, and the roof plains, while the inferential step consists of a fixed derivation sequence of rules that involves the derivation of model parameters. Trying on the one hand to avoid the association of our vocabulary with numerous rules and on the other, aiming at exploiting all the actual design knowledge of our problem, we formulated our vocabulary through hierarchical representations in order to address the problem's complexity while at the same time recover representations of acceptable resolution.

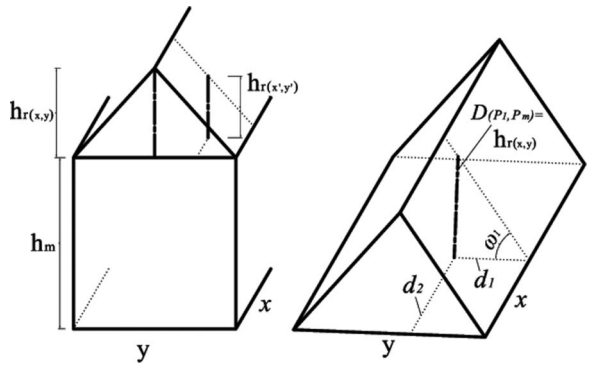

Fig. 3. Hierarchical grammar-based 3-D description for the building models. The building's footprint is determined implicitly from E_{2D}. The building's main height h_m and roof heights $h_r(x, y)$ at every point are recovered (E_{3D}), and thus, all **j** different roof types are modeled or easily derived.

Therefore, the shape-prior formulation involves two components: the type of footprint and the type of roof (Fig. 2). First, we structure our prior-model space $\tilde{\Phi}$ by ascribing the same pointer **i** to all models that belong to the family with the same footprint. Thus, all buildings that can be modeled with a rectangular footprint have the same index value **i**. Then, for every family (i.e., every **i**), the different types of building tops (roofs) are modeled by the pointer **j** (Fig. 2). Under this hierarchy $\tilde{\Phi}_{\mathbf{i},\mathbf{j}}$, the prior database can model from simple to very complex building types and can be easily enriched with more complex structures. Such a formulation is desirously generic but forms a huge search space. Therefore, appropriate attention is to be paid when structuring the search step.

Given the set of footprint priors, we assume that the observed building is a projective transformation of the footprint. Given, the variation of the expressiveness of the grammar and the degrees of freedom of the transformation, we can now focus on the 3-D aspect of the model. In such a context, only the building's main height h_m and the building's roof height $h_r(x, y)$ at every point need to be recovered. The proposed typology for such a task is shown in Fig. 3. It refers to the rectangular case, but all the other families can respectively be defined. More complex footprints, with usually more than one roof type, are decomposed to simpler parts which can, therefore, similarly be recovered. Given an image $\mathcal{I}(x, y)$ at domain (bounded) $\Omega \in R^2$ and an elevation map $\mathcal{H}(x, y)$—which can be seen both as an image or as a triangulated point cloud–let us denote by h_m the main building's height and by P_m the horizontal building's plane at that height. We proceed by modeling all building roofs

(flat, shed, gable, etc.) as a combination of four inclined planes. We denote by P_1, P_2, P_3, and P_4 these four roof planes and by ω_1, ω_2, ω_3, and ω_4, respectively, the four angles between the horizontal plane h_m and each inclined plane (Fig. 3). Every point in the roof rests strictly on one of these inclined planes, and its distance with the horizontal plane is the minimum compared with the ones formed by the other three planes.

With such a grammar-based description, the five unknown parameters to be recovered are as follows: the main height h_m (which has a constant value for every building) and the four angles ω. In this way, all but two types of building tops/roofs can be modeled. For example, if all angles are different, we have a totally dissymmetric roof (Fig. 2—$\tilde{\Phi}_{\mathbf{1,5}}$), if the two opposite angles are right then we have a gable-type one (Fig. 2—$\tilde{\Phi}_{\mathbf{1,4}}$), and if all are zero, we have a flat one ($\tilde{\Phi}_{\mathbf{1,1}}$). The platform and the gambrel roof types cannot be modeled but can be easily derived. The platform one ($\tilde{\Phi}_{\mathbf{1,2}}$), for instance, is the case where all angles have been recovered with small values, and a search around their intersection point will estimate the dimensions of the rectangular-shape box above the main roof plane P_m (e.g., Fig. 10). With the aforementioned formulations, instead of searching for the best among $\mathbf{i} \times \mathbf{j}$ (e.g., $5 \times 6 = 30$) models, their hierarchical grammar and the appropriately defined energy terms are able to cut down effectively the solution space.

More specifically, although such a formulation is desirously generic, it forms a huge search space, and thus, appropriate attention has to be paid when structuring the search step. Toward this end, based on a fruitful synergy with our energy terms (detailed in the following section), we avoid modifying all the possible properties of the basic shapes from our vocabulary, reducing the number of necessary rules. In this way, a nice balance was introduced between the power of the modeling framework (the syntax and semantics of the grammar) and the complexity of the rules required by the framework. This fact is maybe not too obvious for a not-so-extensive vocabulary like the one employed here, but this is due to the resolution of the data that this system is designed to process and the detail of the 3-D geometry that needs to be recovered. A small vocabulary proved sufficient enough. Richer grammars with intermediate structure and shape classification are necessary for more detailed full-scene reconstruction applications fusing data of higher quality from aerial, satellite, or ground sensors [34].

III. Multiple 3-D Building Priors in Competition

Let us consider a pair of images: one that corresponds to the visible domain ($\mathcal{I}$) and the corresponding digital elevation map ($\mathcal{H}$). In such a context, one has first to separate the buildings from the background (natural scene), extract the corresponding footprint types, and determine their geometry. If we consider the two images, then this can be formulated as a segmentation problem. Since neither the number nor the topology of the scene is known, we employ the level-set methods [35]. This can be achieved through the deformation of an initial surface that aims at separating the natural components of the scene from the man-made parts. Let $\phi : \Omega \rightarrow \mathcal{R}^+$ be a level-set representation defined at the dense-image resolution level. We assume that one can establish correspondences between the pixels of the image and the ones of the DEMs. Then, the segmentation can be solved in both spaces ($\mathcal{R}^2$) through the use of regional statistics. In the visible image, we would expect that the buildings are different from the natural components of the scene. On top of that, in the DEM, one would expect that man-made structures will exhibit elevation differences from the natural part of the scene. These two assumptions can be used to define the following segmentation function:

$$\begin{aligned} E_{\text{seg}}(\phi) = & \int |\nabla \phi(\mathbf{x})| \, \mathrm{d}\mathbf{x} + \int_{\Omega} H_\epsilon(\phi) r_{\text{obj}} \left(\mathcal{I}(\mathbf{x})\right) \\ & + \left[1 - H_\epsilon(\phi)\right] r_{\text{bg}} \left(\mathcal{I}(\mathbf{x})\right) \mathrm{d}\mathbf{x} \\ & + \varrho \int_{\Omega} H_\epsilon(\phi) r_{\text{obj}} \left(\mathcal{H}(\mathbf{x})\right) + \left[1 - H_\epsilon(\phi)\right] \\ & \times r_{\text{bg}} \left(\mathcal{H}(\mathbf{x})\right) \mathrm{d}\mathbf{x} \end{aligned} \tag{1}$$

where H is the Heaviside step function, and r_{obj} and r_{bg} are *object* and *background* are positive monotonically decreasing data-driven functions driven from the grouping criteria. The simplest possible approach would involve the Mumford–Shah [36] approach that aims at separating the means between the two classes. In general, choosing the appropriate region descriptors (r_{obj} and r_{bg}) depends heavily on the nature of the images to be considered. One can model the scene in regions with desired objects and in the background and then assume that these regions are characterized by Gaussian densities [37]. When such an assumption seems unrealistic, one can consider a more flexible parametric density function, like a Gaussian mixture [38], or nonparametric densities [39] in order to describe the visual properties of the object and the background. Furthermore, in cases where color information or other remote-sensing data like radar or hyperspectral imagery is available, these region descriptors can be accordingly formulated. One should note the following: 1) that the aim of this paper is not to address the image component of the method and 2) that such image components are defined in a modular content and can easily be adapted to the image content. Therefore, we will assume a rather simple segmentation component just for demonstration purposes. To this end, in all our experiments and in similar manner with [37] and [40], the following region descriptors were employed:

$$r_{\text{obj}} \left(\mathcal{A}(\mathbf{x})\right) = \frac{\left(\mu_{\text{obj}} - \mathcal{A}(\mathbf{x})\right)^2}{\sigma_{\text{obj}}^2} \quad r_{\text{bg}} \left(\mathcal{A}(\mathbf{x})\right) = \frac{\left(\mu_{\text{bg}} - \mathcal{A}(\mathbf{x})\right)^2}{\sigma_{\text{bg}}^2}$$

where $\mathcal{A}$ is either $\mathcal{I}$ or $\mathcal{H}$ and μ_{obj} is the mean and σ_{obj} the covariance matrix of the object appearance (similar definition for the background). Using such a formulation, the scene was modeled as a collection of smooth surfaces and a background, based on observations made at every iteration.

Although, such a data-driven formulation has been considered frequently in computer vision, it is based on a rather simplistic assumption of homogeneity and therefore fails when it is violated. For example, in Fig. 1, the output result from such a purely data-driven term is shown superimposed on

the initial image. The condition of nonhomogeneous regions arises frequently in satellite imaging (both in optical images and DEMs) due to shadows, partial occlusion of the objects of interest, registration errors, poor reflectivity properties, etc. In order to cope with the lack of visual support, one can consider the use of prior knowledge [41]. This can be achieved through the integration of global shape prior constraints into the segmentation process. These constraints can encode both 2-D as well as 3-D measurements. The 2-D constraint, which indicates the type of building footprint, can be determined by fusing our observations, while the 3-D constraint can be determined from the DEM. Let us now consider an abuse of notation and introduce an additional prior component in the process $E_{\text{prior}} = E_{2D} + E_{3D}$.

A. Multiscale Projective-Invariant Footprint Registration

Let us first consider the footprint prior. In order to facilitate the introduction of the concept, we will assume that the building which corresponds to the observed footprint is known. Then, the observed image depends on the pose of the sensor, and therefore, a geometric transformation is to be considered toward establishing a correspondence between the model and the extracted footprint in the image. In the most general case, if $\tilde{\phi}$ is the prior model, then this geometric transformation will minimize the following function [27]:

$$E_{2D}(\phi, \mathcal{T}) = \int_{\Omega} \left(H_\epsilon\left(\phi(\mathbf{x})\right) - H_\epsilon\left(\tilde{\phi}\left(\mathcal{T}(\mathbf{x})\right)\right)\right)^2 \mathrm{d}\mathbf{x} \quad (2)$$

with $\mathcal{T}$ being the admissible geometric relation between the two corresponding shape contours. In the context of our work, we have assumed that a planar projective homography is a reasonable selection. Such a transformation is a mapping $M : \mathcal{P}^2 \rightarrow \mathcal{P}^2$ such that points p_i are collinear if and only if $M(p_i)$ are collinear (projectivity preserves lines) [1].

Following the formulations of [40] and [42], the homograph is calculated directly in its explicit form $\mathcal{T} = r + ((1/d)tn^T)$, where $\mathcal{T}$ is the homography matrix determined by the translation and rotation between the two views t and r and by the structure parameters n and d of the world plane. The translation is described by the vector $t = (t_x, t_y, t_z)$, the rotation matrix $r \in R^3$ (constrained by the three angles α, β, and γ), and, since the world plane is not generally perpendicular to the optical axis, the unit vector n is obtained by first rotating it by an angle ξ around the y-axis and then by an angle ψ around the x-axis. Thus, the nine pose parameters $\mathcal{T}(\alpha, \beta, \gamma, t_x, t_y, t_z, \xi, \psi, d)$ need to be estimated.

Such a function (2) will constrain the segmentation process with respect to a single prior. However, in our case, one has to account for multiple priors. This can be implemented either through a competition approach where all priors are considered and the one performing better is retained or through a vector-valued labeling [40], [43]. This function can be considered to address multiregion segmentation. The role of the labeling function is to evolve dynamically in order to select/indicate the regions where a given prior $\tilde{\phi}_{\mathbf{i}}$ is to be enforced.

For the general case with a large number of building priors $(\tilde{\Phi}_{\mathbf{i},\mathbf{j}})$ and possibly some further independent unknown objects (which should therefore be segmented based on their intensity only), we employed a vector-valued labeling function $\mathbf{L} : \Omega \rightarrow R^k$, $\mathbf{L}(\mathbf{x}) = (L_1(\mathbf{x}), \ldots, L_k(\mathbf{x}))$. The $\nu = 2^k$ vertices of the polytope $[-1, +1]^k$ yield to ν different regions $L_j \in \{+1, -1\}$. The indicator function for each of these regions is denoted by $x_i = 1, \ldots, \nu$. Each indicator function x_i has the form [44]

$$x_i(\mathbf{L}) = \frac{1}{4^k} \prod_{j=1}^{k} (L_j - w_j)^2, \quad \text{with } w_j \in \{-1, +1\}. \quad (3)$$

For example, in cases where $k = 2$, then the indicator function models four regions, i.e.,

$$\begin{aligned} x_1(\mathbf{L}) &= \frac{1}{4^2}(L_1 - 1)^2(L_2 - 1)^2 \\ x_2(\mathbf{L}) &= \frac{1}{4^2}(L_1 + 1)^2(L_2 - 1)^2 \\ x_3(\mathbf{L}) &= \frac{1}{4^2}(L_1 - 1)^2(L_2 + 1)^2 \\ x_4(\mathbf{L}) &= \frac{1}{4^2}(L_1 + 1)^2(L_2 + 1)^2. \end{aligned}$$

With the aforementioned k-dimensional labeling formulation, which is capable of for dynamic labeling of up to $\nu = 2^k$ regions, the following cost function can account for a recognition-driven segmentation, based on multiple competing shape priors

$$\begin{aligned} E_{2D}(\phi, \mathcal{T}_i, \mathbf{L}) = \sum_{i=1}^{\nu-1} \int \left(\frac{H_\epsilon\left(\phi(\mathbf{x})\right) - H_\epsilon\left(\tilde{\phi}_i\left(\mathcal{T}_i(\mathbf{x})\right)\right)}{\sigma_i} \right)^2 \\ \times x_i\left(\mathbf{L}(\mathbf{x})\right) \mathrm{d}\mathbf{x} + \int \lambda^2 x_\nu\left(\mathbf{L}(\mathbf{x})\right) \mathrm{d}\mathbf{x} \\ + \rho \sum_{i=1}^{\nu} \int |\nabla L(\mathbf{x})| \, \mathrm{d}\mathbf{x} \end{aligned} \quad (4)$$

with the two parameters λ and $\rho > 0$. The term associated with the two objects are normalized with respect to the variance of the respective template: $\sigma_i^2 = \int \phi_i^2 \mathrm{d}\mathbf{x} - \int \phi_i \mathrm{d}\mathbf{x}^2$. Contrary to [43], the labeling function's dimensionality k is not *a priori* fixed and is calculated during optimization. Let a positive scalar q denote the number of resulting segments from the image-driven functional. Then

$$k = \lceil \log(1 + q) / \log 2 \rceil .$$

In this way, during optimization, the number of selected regions $\nu = 2^k$ depends on the number of possible building segments according to ϕ, and thus, the k-dimensional labeling function $\mathbf{L}$ obtains incrementally multiple instances. In Figs. 4 and 5, the optimization procedures yield a 2-D labeling. With such a labeling $\nu = 4$, the indicator function models four regions, and for Fig. 5, for example, the first three are responsible for the three detected buildings while the fourth one for the background. It should be also mentioned that here, the initial poses of the priors are not known.

B. Grammar-Based Building Reconstruction

In order to determine the 3-D geometry of the buildings, one has to estimate the height of the structure with respect to the

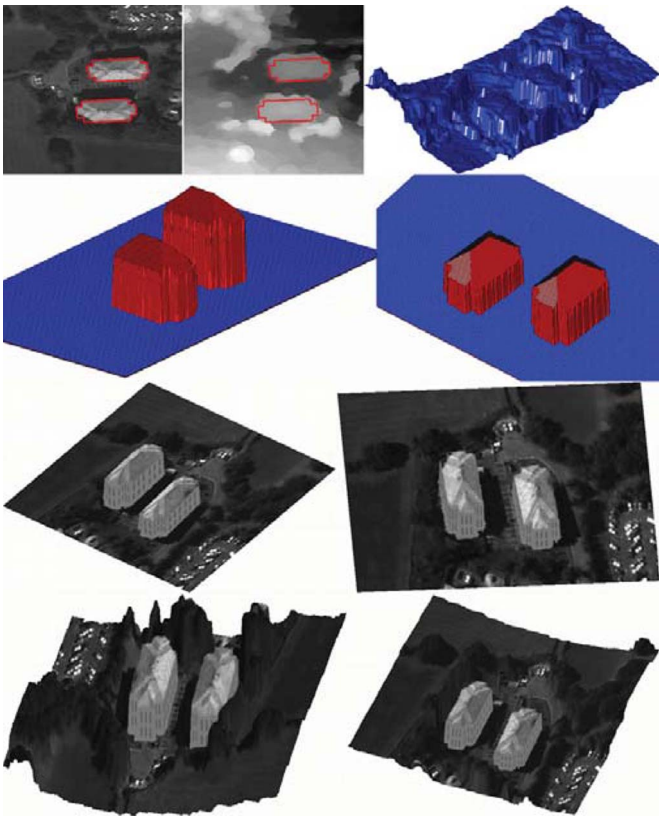

Fig. 4. (First row) Detected building footprints superimposed on the data and a 3-D visualization of the DEM. (Second and third rows) 3-D views of the reconstructed buildings with and without texture. (Fourth row) 3-D views of the scene's reconstruction.

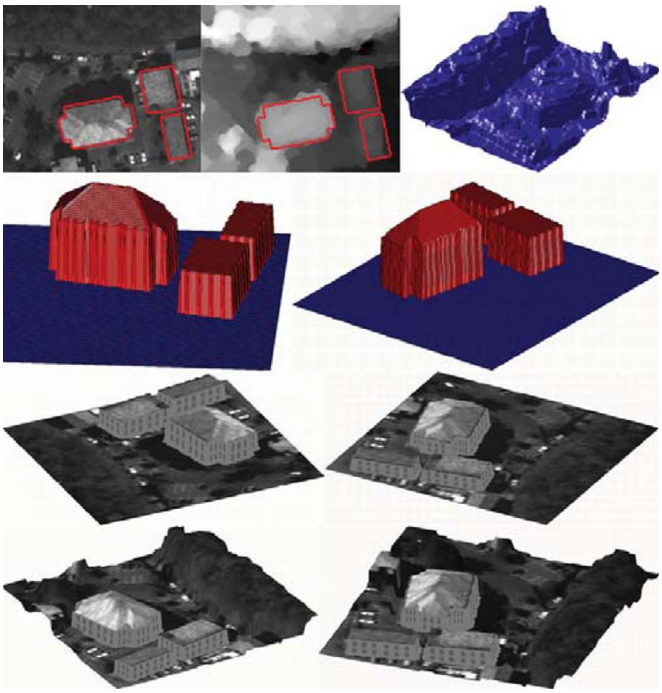

Fig. 5. (First row) Detected building footprints superimposed on the data and a 3-D visualization of the DEM. (Second and third rows) 3-D views of the reconstructed buildings with and without texture. (Fourth row) Reconstructed scene.

ground and the orientation angles of the roof components i.e., five unknown parameters: the building's main height h_m which has a constant value for every building and the four angles ω of the roof's inclined planes $(\Theta_i = (h_m, \omega_1, \omega_2, \omega_3, \omega_4))$. These four angles (Fig. 3), along with the implicitly derived dimensions of every building's footprint (from E_{2D}), can define the roof's height at every point (pixel) $h_r(x, y)$

$$\begin{aligned} &h_r(x,y) \\ &\quad = \min\left[\mathcal{D}(P_1, P_m); \mathcal{D}(P_2, P_m); \mathcal{D}(P_3, P_m); \mathcal{D}(P_4, P_m)\right] \\ &\quad = \min[d_1 \tan\omega_1; d_2 \tan\omega_2; d_3 \tan\omega_3; d_4 \tan\omega_4] \end{aligned} \tag{5}$$

where $\mathcal{D}$: is the perpendicular distance between the horizontal plane P_m and the roof's inclined plane $P_{1:4}$. The distance, e.g., between P_1 and P_m in Fig. 3, is the actual roof's height at that point (x, y) and can be calculated as the product of the tangent of the plane's P_1 angle and the horizontal distance d_1 lying on the plane P_m. $\mathcal{D}(P_1, P_m)$ is also the minimum distance in that specific point compared with the ones that are formed with the other three inclined planes.

Utilizing the 3-D information from $\mathcal{H}$—either from point clouds or from a height map—the corresponding energy E_{3D} that recovers our five unknowns for a certain building i has been formulated as follows:

$$E_{3D}(\Theta_i) = \sum_{i=1}^{m} \int_{\Omega_i} \left(h_{m_i} + h_{r_i}(\mathbf{x}) - \mathcal{H}(\mathbf{x})\right)^2 \mathrm{d}\mathbf{x}. \tag{6}$$

Each prior that has been selected for a specific region is forced to acquire such a geometry so that at every point, its total height matches the one from the available DEM. It is a heavily constrained formulation and, thus, robust. The recognition-driven reconstruction framework introduced here now takes the following form with respect to ϕ, $\mathcal{T}_i$, $\mathbf{L}$, and Θ_i:

$$E_{\text{total}} = E_{\text{seg}}(\phi) + \mu E_{2D}(\phi, \mathcal{T}_i, \mathbf{L}) + \mu E_{3D}(\Theta_i). \tag{7}$$

The energy term E_{seg} addresses fusion in a natural way and solves segmentation ϕ in both the $\mathcal{I}(\mathbf{x})$ and $\mathcal{H}(\mathbf{x})$ spaces. The term E_{2D} estimates which family of priors (i.e., which 2-D footprint $\mathbf{i}$) under any projective transformation T_i best fit at each segment $(\mathbf{L})$. Finally, the energy E_{3D} recovers the 3-D geometry Θ_i of every prior by estimating the building's h_m and h_r heights.

C. Energy Minimization

In order to minimize the energy function (7), one can consider a gradient-descent approach which will update simultaneously ϕ, $\mathcal{T}_i$, $\mathbf{L}$, and Θ_i.

1) Evolution of the Segmentation: For fixed labeling and transformation parameters, the level-set function ϕ evolves according to

$$\begin{aligned} \frac{\vartheta E_{\text{total}}}{\vartheta \phi} &= \frac{\vartheta E_{\text{seg}}}{\vartheta \phi} \\ &\quad - 2\mu \sum_{i=1}^{\nu-1} \frac{H_\epsilon\left(\phi(\mathbf{x})\right) - H_\epsilon\left(\tilde{\phi}_i\left(\mathcal{T}_i(\mathbf{x})\right)\right)}{\sigma_i^2} x_i(\mathbf{L}). \end{aligned} \tag{8}$$

Apart from the first image-driven component, there is an additional relaxation term toward the prior $\tilde{\phi}_{\mathbf{i}}$ in all image regions where $x_i > 0$. Thus, the segmentation favors the curve propagation in regions indicated by the labeling function.

2) Evolution of the k-Dimensional Labeling Function: For a fixed level-set function ϕ and transformation parameters, the gradient descent with respect to the labeling functions L_i corresponds to an evolution of the form

$$\frac{\vartheta L_j}{\vartheta t} = -\mu \sum_{i=1}^{\nu-1} \frac{\left(H_\epsilon\left(\phi(\mathbf{x})\right) - H_\epsilon\left(\tilde{\phi}_i\left(\mathcal{T}_i(\mathbf{x})\right)\right)\right)^2}{\sigma_i^2} \frac{\vartheta x_i}{\vartheta L_j} - \mu\, \lambda^2 \frac{\vartheta x_\nu}{\vartheta L_j} - \mu\, \rho\, div \frac{\nabla L_j}{\|\nabla L_j\|} \quad (9)$$

where the derivatives of the indicator functions x_i are calculated from (3). The first two terms guide the labeling $\mathbf{L}$ to indicate the transformed priors $\tilde{\phi}_{\mathbf{i}}$ which are most similar to the given function ϕ (i.e., each labeled segment or the background). The last term imposes spatial regularity in the labeling L_j and enforces the selected regions to be compact by preventing flippings with the neighboring locations.

3) Pose Estimation: The optimization of the projective transformation parameters $\mathcal{T}(\alpha_i, \beta_i, \gamma_i, (t_x)_i, (t_y)_i, (t_z)_i, \xi_i, \psi_i, d_i)$ of each selected prior $\tilde{\phi}_{\mathbf{i}}$ was achieved with a multiscale process, in order to handle both global and local shape deformations. The multiscale approach is implemented via a fixed-point iteration on both the level-set function ϕ and shape priors $\tilde{\phi}_{\mathbf{i}}$ with a downsampling strategy by a factor l. The general gradient-descent equation for each of the transformation parameters (denoted by u_i) has, thus, the following form:

$$\frac{\vartheta u_i^l}{\vartheta t} = 2\mu x_i^l(\mathbf{L}^l) \int_\Omega \left(\frac{H_\epsilon\left(\phi^l(\mathbf{x})\right) - H_\epsilon\left(\tilde{\phi}_i^l\left(\mathcal{T}_i^l(\mathbf{x})\right)\right)}{\sigma_i^2} \right) \times \frac{\vartheta \mathcal{T}_i^l\left(u_i^l\right)}{\vartheta u_i^l}. \quad (10)$$

4) Evolution of the Building's 3-D Geometry: For a fixed segmentation ϕ, a labeling L, and transformation parameters $\mathcal{T}_i$, the 3-D geometry Θ_i of each selected building model is derived by a gradient-descent process with respect to the building's height h_m and the four angles $\omega_\eta, \eta\{1:4\}$. Computing $\vartheta E_{3D}/\vartheta h_m$ is straightforward, while

$$\frac{\vartheta E_{3D}}{\vartheta \omega_\eta} = 2\left(h_m + h_r(x,y) - \mathcal{H}(x,y)\right) \frac{\vartheta h_r(x,y)}{\vartheta \omega_\eta} \quad (11)$$

where

$$\frac{\vartheta h_r(x,y)}{\vartheta \omega_\eta} = \frac{\vartheta \min[d_1 \tan\omega_1; d_2 \tan\omega_2; d_3 \tan\omega_3; d_4 \tan\omega_4]}{\vartheta \omega_\eta} \quad (12)$$

which can be calculated with the following rules:

$$\min(x1; x2) = 0.5\left(x1 + x2 - \sqrt{(x1 - x2)^2}\right)$$

$$\min(x1; x2; x3; x4) = \min\left(\min(x1; x2); \min(x3; x4)\right).$$

IV. Evaluation

A. Quantitative Measures

The quality assessment of 3-D data ([45], [46] and their references therein) involves the assessment of both the geometry and topology of the model. During our experiments, the quantitative evaluation was performed based on the 3-D ground-truth data which were derived from a manual digitization procedure. The standard quantitative measures of Completeness, Correctness, and Quality (a normalization between the previous two) were employed. To this end, the quantitative assessment is divided into two parts: First, for the evaluation of the extracted 2-D boundaries, i.e., the horizontal localization of the building footprints (like those shown in Fig. 11) and, second, for the evaluation of the hypsometric differences, i.e., the vertical differences between the extracted 3-D building and the ground truth (like those shown in Fig. 12).

In order to assess the horizontal accuracy of the extracted building footprints, the measures of horizontal true positives (HTPs), horizontal false positives (HFPs), and horizontal false negatives (HFNs) were calculated, i.e.,

$$\text{2D Completeness} = \frac{\text{area of correctly detected segments}}{\text{area of the ground truth}} = \frac{\text{HTP}}{\text{HTP} + \text{HFN}}$$

$$\text{2D Correctness} = \frac{\text{area of correctly detected segments}}{\text{area of all detected segments}} = \frac{\text{HTP}}{\text{HTP} + \text{HFP}}$$

$$\text{2D Quality} = \frac{\text{HTP}}{\text{HTP} + \text{HFP} + \text{HFN}}.$$

Moreover, for the evaluation of the hypsometric differences between the extracted buildings and the ground truth, the measures of vertical true positives (VTPs), vertical false positives (VFPs), and vertical false negatives (VFNs) were also calculated. The VTP are the voxels among the corresponding HTP pixels that have the same altitude with the ground truth. Note that HTPs may correspond to the following: 1) to voxels with the same altitude as in the ground truth (VTP) and 2) to voxels with a lower or higher altitude than the ground truth (VFN and VFP, respectively). Thus, the VFPs are the voxels with a hypsometric difference with the ground truth, containing all the corresponding voxels from the HFP and the corresponding ones from the HTP (those with a higher altitude than the ground truth). Respectively, the VFNs are the voxels with a hypsometric difference with the ground truth, containing all the corresponding voxels from the HFN and the corresponding ones from the HTP (those with a lower altitude than the ground truth). To this end, the 3-D quantitative assessment was based

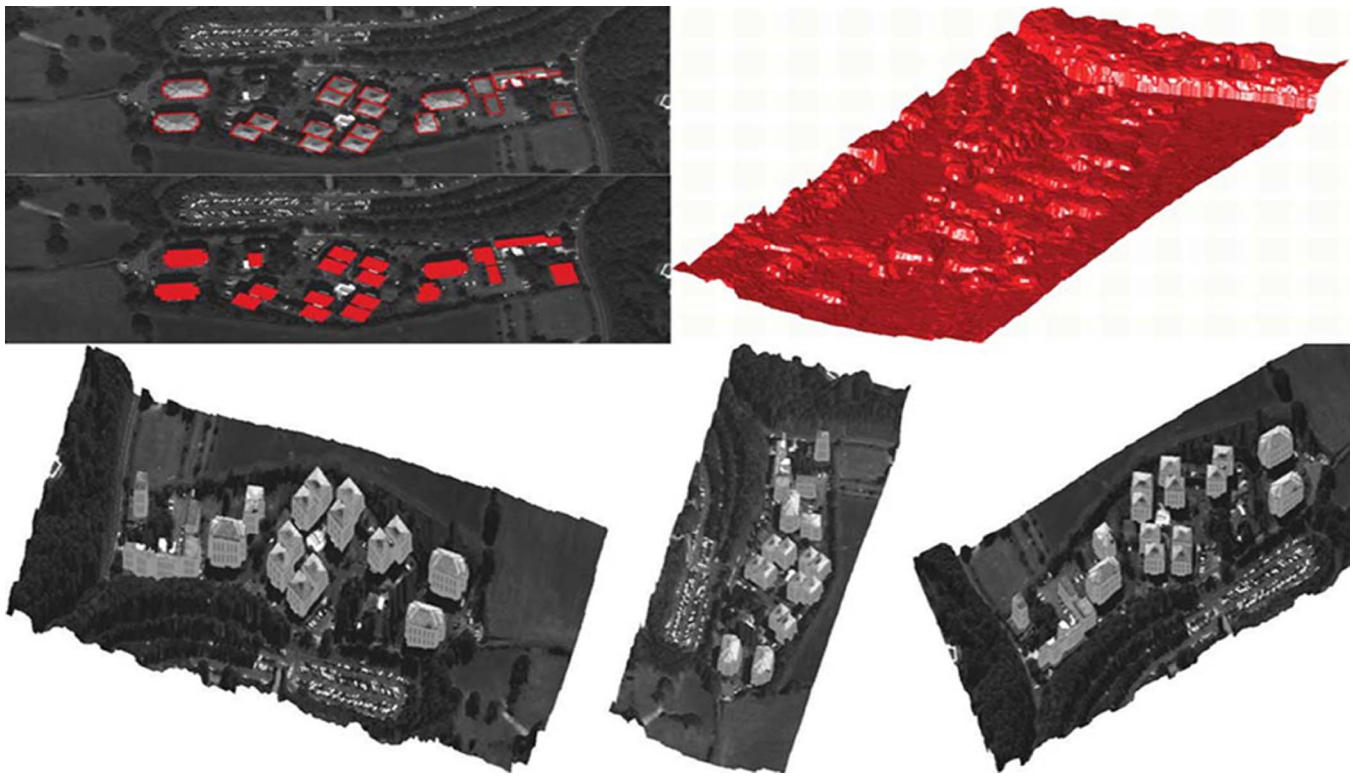

Fig. 6. Developed algorithm can account for important terrain's height variability and overcome detection errors due to shadows, occlusions, and conflicting similar neighboring heights. (First row top) Detected building boundaries superimposed on the initial image. (First row bottom) Ground truth superimposed on the initial image. (First row right) 3-D view of scene's DEM. (Bottom) 3-D views of the reconstructed scene.

on the measures of the 3-D Completeness, 3-D Correctness, and 3-D Quality (a normalization between the previous two), which were calculated in the following way:

$$\text{3D Completeness} = \frac{\text{VTP}}{\text{VTP} + \text{VFN}}$$

$$\text{3D Correctness} = \frac{\text{VTP}}{\text{VTP} + \text{VFP}}$$

$$\text{3D Quality} = \frac{\text{VTP}}{\text{VTP} + \text{VFP} + \text{VFN}}.$$

B. Experimental Results

The developed algorithm has been applied to a number of scenes where remote-sensing data were available.[1] In Figs. 4 and 5, the results for the detection and reconstruction of a small number of buildings are presented. The algorithm managed in all cases to accurately recover their footprint and overcome low-level misleading information due to shadows, occlusions, etc. In addition, despite the conflicting height similarity between the desired buildings, the surrounding trees, and the other objects, the developed algorithm managed to robustly recover their 3-D geometry as the appropriate priors were chosen. In both cases of Figs. 4 and 5, the performed quantitative evaluation indicated that the algorithm's completeness, correctness, and overall quality-standard quantitative measures for man-made object extraction was 98% and 96%, respectively.

[1] http://www.mas.ecp.fr/vision/Personnel/karank/Demos/3D

TABLE I
PIXEL- AND VOXEL-BASED QUALITY ASSESSMENT

2D Quantitative Measures			
	Completeness	Correctness	Quality
Dataset #1	0.88	0.93	0.82
Dataset #2	0.87	0.98	0.85
Dataset #3	0.84	0.90	0.76

3D Quantitative Measures			
	Completeness	Correctness	Quality
Dataset #1	0.86	0.93	0.80
Dataset #2	0.87	0.95	0.83
Dataset #3	0.86	0.86	0.77

In Fig. 6, results are shown for a quite complex scenario (data set #1). The considered areas consist of complex landscape, multiple objects of various classes, shadows, occlusions, different texture patterns, and an important terrain variability. For both test site, just a single panchromatic aerial image with approximately 0.7-m spatial resolution and the corresponding DEM in a lower resolution (of approximately 2.5 m) were available. The detected building footprints superimposed on the data are shown in Fig. 6 (left) and a 3-D view of the recovered 3-D geometry is shown in Fig. 6 (right). All buildings—except the one at the top left of the scene—were extracted and reconstructed. All of them have been recognized with a different

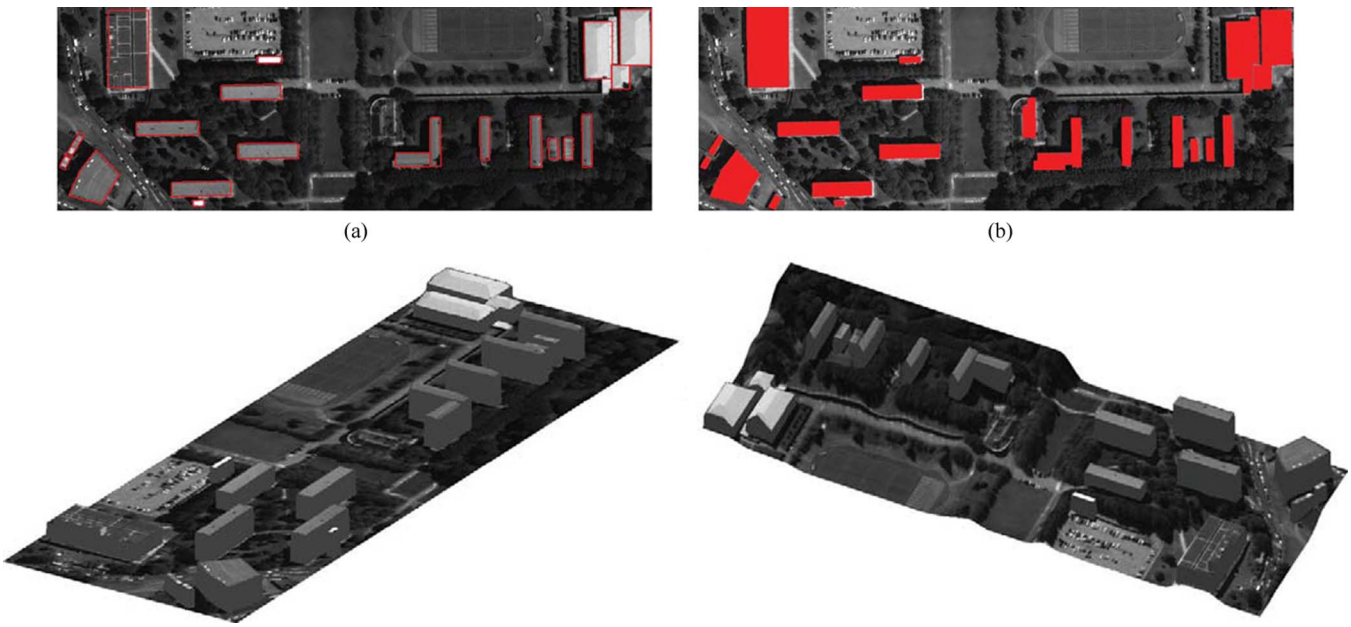

Fig. 7. Large-scale building reconstruction. (Left) A 3-D view of the reconstructed buildings and (right) a 3-D view of the entire scene's reconstruction. (a) Detected building boundaries. (b) Ground truth.

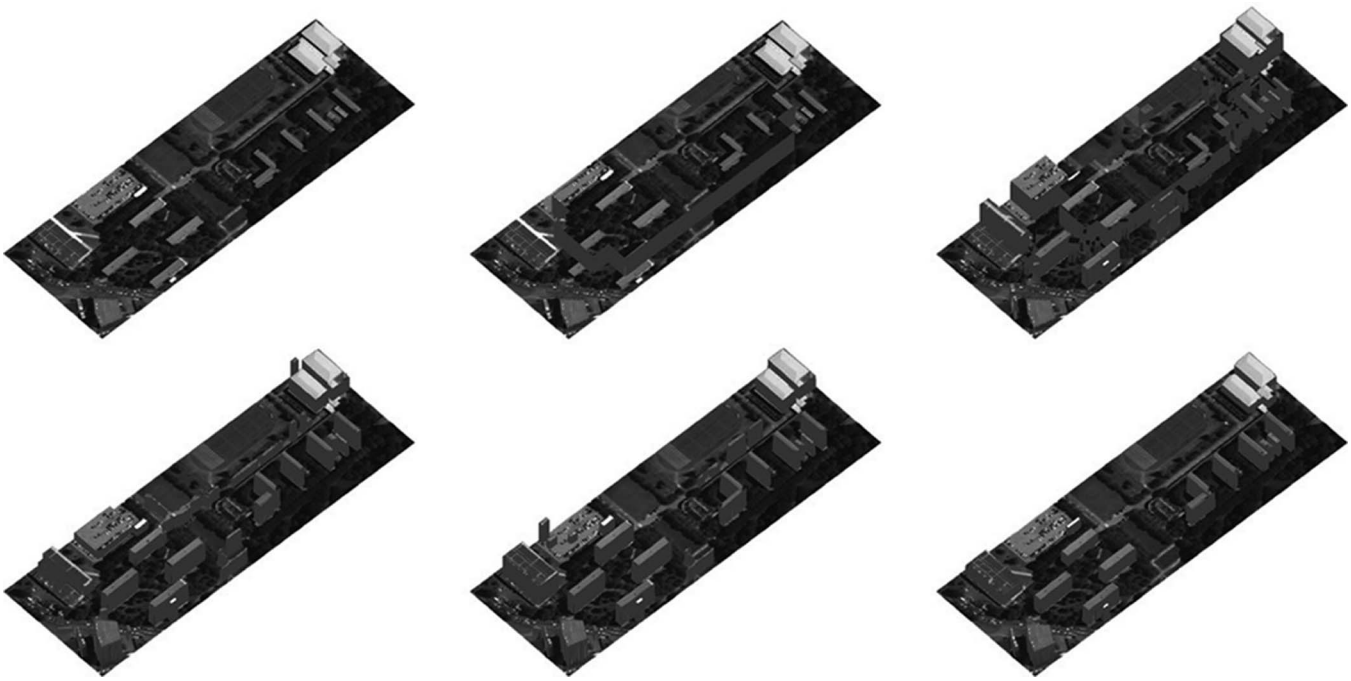

Fig. 8. Different optimization steps toward the scene's reconstruction. (Top left) Starting from the initial image, after a small number of iterations, (bottom right) the algorithm converged and managed to accurately recover the scene's 3-D geometry.

identity (have been labeled and numbered uniquely) apart from the three-building segment at the top right corner of the scene. It was poorly detected but also appears as one segment in the ground-truth data. The performed quantitative evaluation reported an overall horizontal-detection correctness of approximately 93% and completeness of approximately 88%, indicating the algorithm's high potentials. Furthermore, regarding the vertical detection accuracy, the algorithm had an overall 3-D voxel-based detection quality of 80%, with a 3-D completeness of 86% and a higher 3-D correctness of 93% (Table I, data set #1).

Furthermore, the developed algorithm was applied to another test site (data set #2 of similar quality), with important terrain variability and complex landscape, including multiple objects of various classes, shadows, occlusions, and different texture patterns (Fig. 7). The different steps from the optimization procedure are shown in Fig. 8. After a small number of iterations, the algorithm converged and managed to accurately

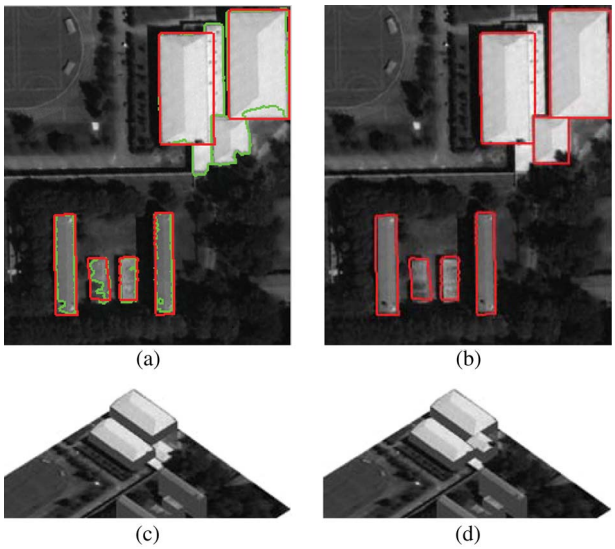

Fig. 9. (First row) (a) At iteration $t-1$, the result of the segmentation energy term (in green color) much differs from the recovered model (in red color), and therefore, the algorithm (with an unsupervised manner) is forced to decompose the registration process. (Second row) 3-D views from the corresponding algorithm's reconstruction results. (a) Detection at iteration: $t-1$. (b) Final detection at iteration t. (c) Recovered geometry at iteration $t-1$. (d) Reconstruction at iteration t.

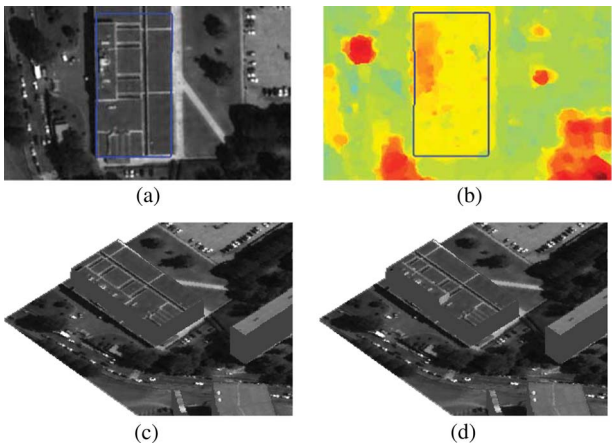

Fig. 10. (a) Detected building boundaries and (b) the corresponding height variation are shown for iteration $t-1$. (c) Temporarily, the roof was wrongly recovered as a flat one, and due to the calculated difference between the recovered model and the height data, the algorithm (with an unsupervised manner) decomposed the reconstruction process. (d) The roof was correctly reconstructed at the next iteration. (a) Detected building boundaries (iteration $t-1$). (b) Height variation inside the detected footprint. (c) The roof is temporarily wrongly recovered as a flat one (iteration $t-1$). (d) Correct reconstruction at iteration t.

recover the scene's 3-D geometry. All buildings–except the one in the middle of the scene–were extracted and accurately reconstructed. Note that in its convergence, the algorithm managed to effectively account for all the roof types. In cases where, during the optimization procedure, there was a significant difference between the calculated inner energies (between data, detection, and the temporal recovered geometry), the algorithm (with an unsupervised manner) was forced to search for an optimal solution. In Figs. 9 and 10, two cases of decomposing the reconstruction task at possible local minima are shown. In Fig. 9, the resulting temporal (iteration $t-1$) detected footprints (in red) that are much different from the ones of the data-driven segmentation term (in green color). Therefore, the algorithm automatically was forced to decompose the registration process seeking for a more optimal solution, which was derived in the next iteration [i.e., comparing the results in red contours of Fig. 9(a) and (b)]. Similarly, in Fig. 10, the roof of the building was wrongly recovered (at iteration $t-1$) as a flat one [Fig. 10(c)]. The calculated difference between the recovered model and the height data [Fig. 10(b)] forced the algorithm to decompose the reconstruction process. The roof was correctly reconstructed at the next iteration [Fig. 10(d)].

The performed quantitative evaluation for the second data set reported an overall detection 2-D correctness of approximately 98%, indicating the algorithm's high potentials. The algorithm's overall detection 2-D completeness was measured at approximately 87%, and its overall detection quality is 85% (Table I, data set #2). The nicely reconstructed buildings and the reconstructed scene are shown in Fig. 7. In Table I, the reported voxel-based evaluation indicated the vertical differences between the ground truth and the reconstructed buildings with a 3-D correctness of 95% and a 3-D completeness of 87%.

Last but not least, the robustness and operational functionality of the proposed method is shown in Fig. 11, where another test site has been reconstructed (data set #3). This complex landscape contains a big variety of texture patterns, more than 80 buildings of different types (detached single-family houses, industrial buildings, etc.) and multiple other objects of various classes. One can directly observe (Fig. 11, first row) that shadows and occlusions were strongly present both on the two available aerial images (with a ground resolution of approximately 0.5 m) and on the coarser digital surface model (of approximately 1.0-m ground resolution). The proposed generic variational framework managed to accurately extract the 3-D geometry of the scene's buildings, searching among various footprint shapes and various roof types. The robustness and functionality of the proposed method is also shown in the second and third row of Fig. 11, where one can clearly observe the HTP, HTN, and HFN. The performed quantitative evaluation reported an overall horizontal detection 2-D correctness of 90% and an overall horizontal detection 2-D completeness of 84% (Table I, data set #3). Compared with the purely image-driven detection results (with an overall detection quality of lower than 65%), the proposed competing 3-D shape priors under a fruitful synergy with the energy terms were able to successfully recover the scene's geometry. In the last row of Fig. 11, the reconstructed buildings and the reconstructed scene are shown, demonstrating that the proposed generic variational framework managed to accurately extract the 3-D geometry of the scene's buildings, searching among various footprint shapes and various roof types.

The aforementioned qualitative observations are supported by the quantitative measures reported in Table I and shown in Fig. 12. More specifically, in Fig. 12, the hypsometric/vertical difference between the extracted buildings and the ground truth is shown. The VFN voxels are in red color, while the VFP ones are in green. Similarly, in Fig. 4(c) the VFN and VFP voxels—corresponding to the HTP pixels—are shown. The performed quantitative evaluation reported both overall 3-D

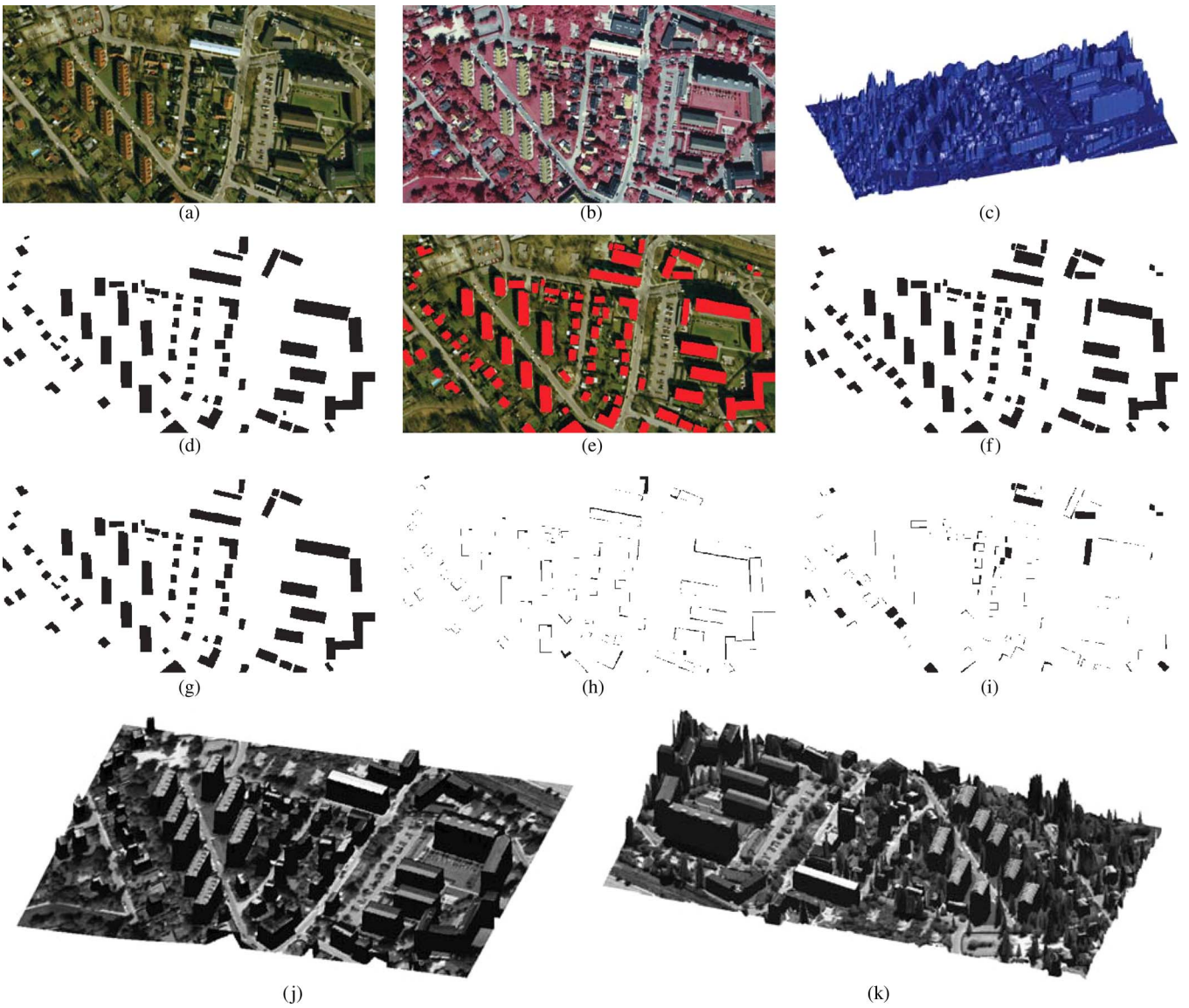

Fig. 11. Large-scale building reconstruction through competing 3-D priors. (a) RGB satellite image. (b) Near-infrared satellite image. (c) 3-D view of the scene's DEM. (d) Detected buildings. (e) Ground truth superimposed on (a). (f) Binary ground truth. (g) HTPs. (h) HFPs. (i) HFNs. (j) Extracted buildings. (k) Reconstructed scene.

completeness and correctness of approximately 86% (Table I, data set #3).

C. Discussion

As has been demonstrated in the experimental results, multiple buildings can be automatically extracted and reconstructed, without any *a priori* information, for their exact shape or number. The selected shape priors geometrically evolve in time (Fig. 8) and determine the 3-D vector description of all the scene's buildings. Although, in this way, a recognition process has been elegantly integrated into a variational segmentation framework, in cases where the data term cannot detect possible building regions, the algorithm naturally fails since all energy terms are associated with the ϕ function. The evolution of the labeling function is driven by the competing shape priors, and each selected image region is ascribed to the best fitted one. The joint multiscale optimization of the transformation parameters allowed keeping track of the correct pose of each object. The function is also consistent with the philosophy of level sets as it allows multiple independent-object detection.

Toward designing a generic framework for automatic 3-D building extraction, in all our experiments, the tuning parameters ϱ, μ, λ, and ρ were left constant, and the texture on the building walls has been added for visualization purposes. In particular, the parameter ϱ (1), which controls which observation affects more the data-driven segmentation term, was set to $\varrho = 0.85$. This was mainly because we were focusing on cases where the higher spatial resolution was on the optical data and not on the DEM, and thus, the segmentation term was constrained accordingly. Regarding the positive weight μ of the shape prior terms (7), it was set empirically to $\mu = 1$ equalizing the importance of all energy terms. Furthermore, the two parameters λ and ρ of the prior term E_{2D} (4) were set to 0.95 and 1, respectively. The parameter $\rho = 1$ acted as a TV regularization operator and forced the boundary that

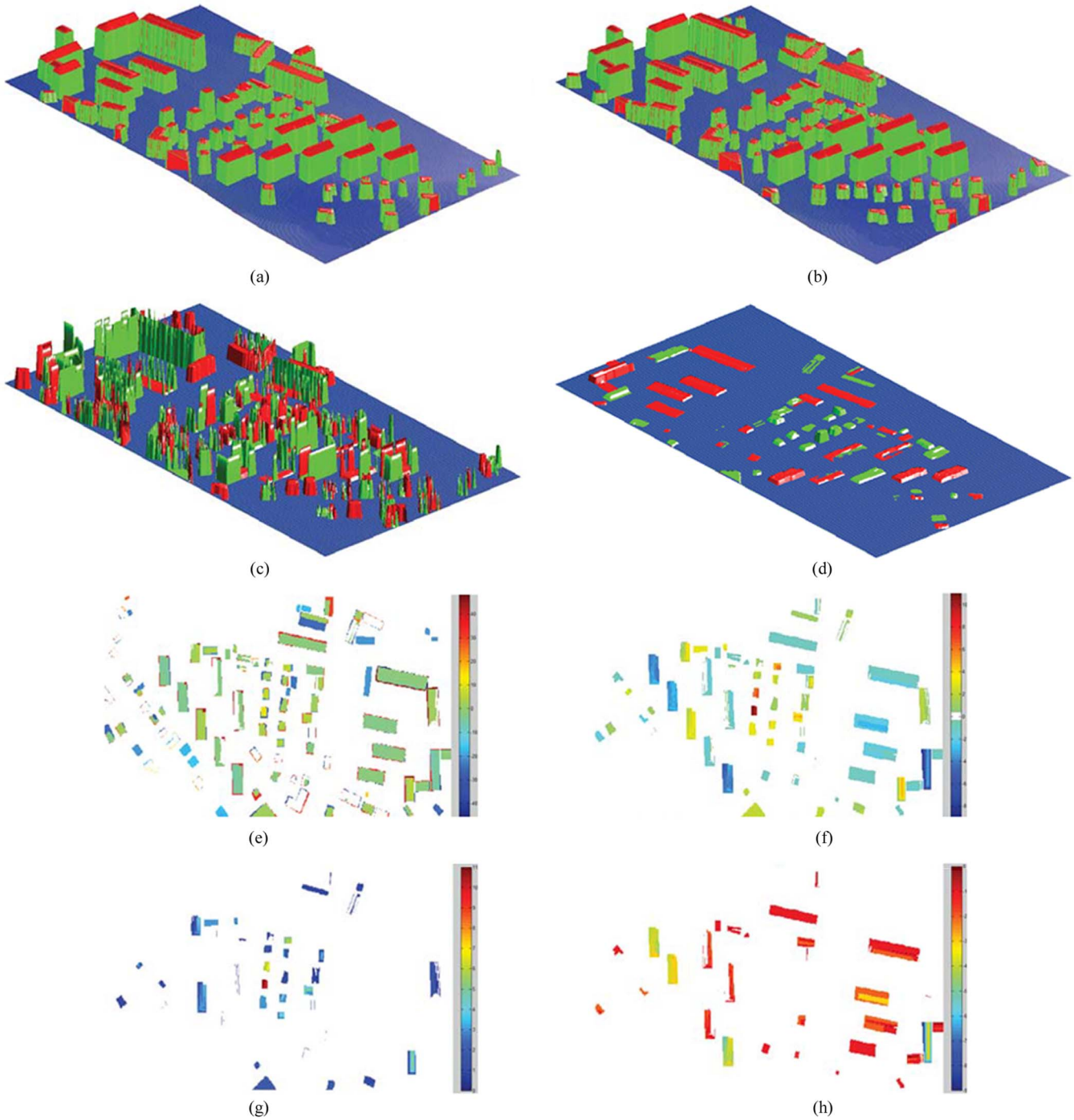

Fig. 12. Vertical/hypsometric difference between the extracted buildings and the ground truth. (a) 3-D view of the extracted buildings. (b) 3-D view of the ground truth. (c) Vertical/Hypsometric difference (absolute values). (d) Vertical difference among the HTPs (absolute values). (e) Vertical difference. (f) Vertical difference among the HTPs. (g) VFPs among the HTPs. (h) VFNs among the HTPs.

separates the labeling regions to have minimal length, imposing spatial regularity in the labeling enforcing the selected regions to be compact by preventing flippings with the neighboring locations. Setting the parameter λ—which balances the competition between the background region and the one of the shape prior—to 0.95 slightly affected the outcome of the segmentation process by decreasing the relative size of the identified background region. Last but not least, we would like to mention that the sensitivity of the proposed framework to the initialization of level-set function was significantly low. The obtained results were practically independent of the initialization.

Furthermore, in all our experiments, the eight (i) by six (j) prior building models shown in Fig. 2 were used, but this database can be updated with other more complex shapes. In cases where the detected building cannot be sufficiently described from any shape from the database—under any possible planar projectivity–the algorithm fails to accurately detect its

boundaries or 3-D geometry. A certain solution is to construct a large database with all the representative shape samples (derived, for example, from cadastral maps) but then, the computation time will increase a lot. Searching, in our experiments, in a space of eight possible solutions for every detected segment, the developed algorithm in MATLAB, without an optimized coding, managed to converge approximately after a couple of hours (8 to 12 iterations) in an ordinary iPentiumM 2 GHz and for an image of approximately half a million pixels. However, with an efficient C++ implementation, the processing time will be decreased by a factor of 1000, given prior experience in similar problems, allowing near real-time applications. To the best of our knowledge, formulating large scale 3-D reconstruction under a single generic variational framework using such a grammar-based modeling (numerous 3-D competing priors formulated under a narrow search space) and fusion of high and very high resolution images and depth maps, was not done before. This combined grammar-based approach for segmentation and reconstruction can determine accurately (overcoming shadows, occlusions, etc.) the buildings' pose and 3-D geometry based on just a single panchromatic image and an elevation map.

V. Conclusion and Future Perspectives

We have developed a generalized variational framework which addresses large-scale reconstruction through information fusion and competing grammar-based 3-D priors. We have argued that our inferential approach significantly extends previous 3-D extraction and reconstruction efforts by accounting for shadows, occlusions, and other unfavorable conditions, and by effectively narrowing the space of solutions due to our novel grammar representation and energy formulation. The successful recognition-driven results along with the reliable estimation of buildings 3-D geometry suggest that the proposed method constitutes a highly promising tool for various object extraction and reconstruction tasks.

Our framework can be easily extended to process spectral information, by formulating, respectively, the region descriptors and to account for other types of buildings or other terrain features. Real-time applications can be considered through the direct implementation of the level-set evolution function at the level of the graphics processing units (GPUs) that could drastically decrease computational complexity. The proposed method alternates between the estimation of the grammar parameters and the evolution of the level set that does perform the segmentation in the image space. The computational complexity of the method is due to the level-set evolution (99% of the time). The level-set formulation results on individual evolution equations at the pixel level being connected only with the neighboring ones, therefore, a GPU architecture is perfectly suitable for this task. Furthermore, in order to address the suboptimality of the obtained solution, the use of the compressed sensing framework by collecting a comparably small number of measurements rather than all pixel values is currently under investigation. Last but not least, introducing hierarchical procedural grammars [34] can reduce the complexity of the prior model and provide access to more efficient means of optimization.

References

[1] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[2] B. C. Matei, H. Sawhney, S. Samarasekera, J. Kim, and R. Kumar, "Building segmentation for densely built urban regions using aerial LIDAR data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[3] D. Gallup, J. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys, "Real-time plane-sweeping stereo with multiple sweeping directions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[4] P. Gamba and B. Houshmand, "Digital surface models and building extraction: A comparison of IFSAR and LIDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 4, pp. 1959–1968, Jul. 2000.

[5] O. Firschein and T. Strat, *Radius: Image Understanding for Imagery Intelligence*. San Mateo, CA: Morgan Kaufmann, 1997.

[6] A. Gruen and R. Nevatia, Eds., "Automatic building extraction from aerial images," *Comput. Vis. Image Underst.*, vol. 72, no. 2, pp. 99–100, Nov. 1998.

[7] C. Baillard, C. Schmid, A. Zisserman, and A. Fitzgibbon, "Automatic line matching and 3D reconstruction of buildings from multiple views," in *Proc. ISPRS Conf. Autom. Extraction GIS Objects From Digital Imagery*, 1999, vol. 32, pp. 69–80.

[8] J. Hu, S. You, and U. Neumann, "Approaches to large-scale urban modeling," *IEEE Comput. Graph. Appl.*, vol. 23, no. 6, pp. 62–69, Nov./Dec. 2003.

[9] G. Zhou, C. Song, J. Simmers, and P. Cheng, "Urban 3D GIS from LiDAR and digital aerial images," *Comput. Geosci.*, vol. 30, no. 4, pp. 345–353, May 2004.

[10] L. Chen, T. Teo, J. Rau, J. Liu, and W. Hsu, "Building reconstruction from LIDAR data and aerial imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2005, vol. 4, pp. 2846–2849.

[11] H. You and S. Zhang, "3D building reconstruction from aerial CCD image and sparse laser sample data," *Opt. Lasers Eng.*, vol. 44, no. 6, pp. 555–566, Jun. 2006.

[12] A. Zakhor and C. Frueh, "Automatic 3D modeling of cities with multimodal air and ground sensors," in *Multimodal Surveillance, Sensors, Algorithms and Systems*, Z. Zhu and T. Huang, Eds. Norwood, MA: Artech House, 2007, ch. 15, pp. 339–362.

[13] K. Karantzalos and D. Argialas, "A region-based level set segmentation for automatic detection of man-made objects from aerial and satellite images," *Photogramm. Eng. Remote Sens.*, vol. 75, no. 6, pp. 667–678, 2009.

[14] C. Brenner, "Building reconstruction from images and laser scanning," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 6, no. 3/4, pp. 187–198, Mar. 2005.

[15] Z. Zhu and T. Kanade, Eds., "Special issue: Modeling and representations of large-scale 3D scenes," *Int. J. Comput. Vis.*, vol. 78, no. 2/3, pp. 119–120, Jul. 2008.

[16] C. Jaynes, E. Riseman, and A. Hanson, "Recognition and reconstruction of buildings from multiple aerial images," *Comput. Vis. Image Underst.*, vol. 90, no. 1, pp. 68–98, Apr. 2003.

[17] I. Suveg and G. Vosselman, "Reconstruction of 3D building models from aerial images and maps," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, no. 3/4, pp. 202–224, Jan. 2004.

[18] A. R. Dick, P. H. S. Torr, and R. Cipolla, "Modelling and interpretation of architecture from several images," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 111–134, Nov. 2004.

[19] Z. Kim and R. Nevatia, "Automatic description of complex buildings from multiple images," *Comput. Vis. Image Underst.*, vol. 96, no. 1, pp. 60–95, Oct. 2004.

[20] M. Wilczkowiak, P. Sturm, and E. Boyer, "Using geometric constraints through parallelepipeds for calibration and 3D modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 194–207, Feb. 2005.

[21] V. Verma, R. Kumar, and S. Hsu, "3D building detection and modeling from aerial LIDAR data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 2213–2220.

[22] G. Forlani, C. Nardinocchi, M. Scaioni, and P. Zingaretti, "Complete classification of raw LIDAR data and 3D reconstruction of buildings," *Pattern Anal. Appl.*, vol. 8, no. 4, pp. 357–374, Feb. 2006.

[23] F. Lafarge, X. Descombes, J. Zerubia, and M. Pierrot-Deseilligny, "3D city modeling based on hidden Markov model," in *Proc. IEEE ICIP*, 2007, vol. II, pp. 521–524.

[24] F. Rottensteiner, J. Trinder, S. Clode, and K. Kubik, "Building detection by fusion of airborne laser scanner data and multi-spectral images: Performance evaluation and sensitivity analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 2, pp. 135–149, Jun. 2007.

[25] G. Sohn and I. Dowman, "Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 1, pp. 43–63, May 2007.
[26] L. Zebedin, J. Bauer, K. Karner, and H. Bischof, "Fusion of feature- and area-based information for urban buildings modeling from aerial imagery," in *Proc. Eur. Conf. Comput. Vis.*, vol. 5305, *Lecture Notes in Computer Science*, 2008, pp. 873–886.
[27] N. Paragios, Y. Chen, and O. Faugeras, *Handbook of Mathematical Models of Computer Vision*. New York: Springer-Verlag, 2005.
[28] M. Taron, N. Paragios, and M.-P. Jolly, "Registration with uncertainties and statistical modeling of shapes with variable metric kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 99–113, Jan. 2009.
[29] P. Müller, G. Zeng, P. Wonka, and L. Gool, "Image-based procedural modeling of facades," *Proc. ACM SIGGRAPH/ACM Trans. Graph.*, vol. 26, no. 3, 9 pp., 2007.
[30] P. Muller, P. Wonka, S. Haegler, A. Ulmer, and L. Gool, "Procedural modeling of buildings," *Proc. ACM SIGGRAPH/ACM Trans. Graph.*, vol. 25, no. 3, pp. 614–623, 2006.
[31] N. Ripperda and C. Brenner, "Data driven rule proposal for grammar based facade reconstruction," in *Proc. PIA. Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, U. Stilla, H. Mayer, F. Rottensteiner, C. Heipke, and S. Hinz, Eds., 2007, vol. 36, pp. 1–6.
[32] D. Doerschlag, G. Groeger, and L. Pluemer, "Semantically enhanced prototypes for building reconstruction," in *Proc. PIA. Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, U. Stilla, H. Mayer, F. Rottensteiner, C. Heipke, and S. Hinz, Eds., 2007, vol. 36, pp. 111–116. Part 3/W49A.
[33] S. Becker and N. Haala, "Grammar supported facade reconstruction from mobile LIDAR mapping," in *Proc. CMRT. Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, U. Stilla, F. Rottensteiner, N. Paparoditis, Eds., 2007, vol. 38, pp. 229–234.
[34] P. Koutsourakis, L. Simon, O. Teboul, G. Tziritas, and N. Paragios, "Single view reconstruction using shape grammars for urban environments," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009. [Online]. Available: http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?nrf=true&punumber=1000149
[35] S. Osher and N. Paragios, *Geometric Level Set Methods in Imaging, Vision and Graphics*. New York: Springer-Verlag, 2003.
[36] D. Mumford and J. Shah, "Optimal approximation by piecewise smooth functions and associated variational problems," *Commun. Pure Appl. Math.*, vol. 42, no. 5, pp. 577–685, 1989.
[37] T. Chan and L. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, Feb. 2001.
[38] N. Paragios and R. Deriche, "Geodesic active regions: A new framework to deal with frame partition problems in computer vision," *J. Vis. Commun. Image Represent.*, vol. 13, no. 1/2, pp. 249–268, Mar. 2002.
[39] J. Kim, J. Fisher, A. Yezzi, M. Cetin, and A. Willsky, "A nonparametric statistical method for image segmentation using information theory and curve evolution," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1486–1502, Oct. 2005.
[40] K. Karantzalos and N. Paragios, "Recognition-driven 2D competing priors towards automatic and accurate building detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 1, pp. 133–144, Jan. 2009.
[41] M. Rousson and N. Paragios, "Prior knowledge, level set representations and visual grouping," *Int. J. Comput. Vis.*, vol. 76, no. 3, pp. 231–243, Mar. 2008.
[42] T. Riklin-Raviv, N. Kiryati, and N. Sochen, "Prior-based segmentation and shape registration in the presence of perspective distortion," *Int. J. Comput. Vis.*, vol. 72, no. 3, pp. 309–328, May 2007.
[43] D. Cremers, N. Sochen, and C. Schnörr, "A multiphase dynamic labeling model for variational recognition-driven image segmentation," *Int. J. Comput. Vis.*, vol. 66, no. 1, pp. 67–81, Jan. 2006.
[44] T. Chan and W. Zhu, "Level set based shape prior segmentation," Comput. Appl. Math., UCLA, Los Angeles, CA, Tech. Rep. 03-66, 2003.
[45] J. Meidow and H. Schuster, "Voxel-based quality evaluation of photogrammetric building acquisitions," in *Proc. ISPRS Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, U. Stilla, F. Rottensteiner, and S. Hinz, Eds., 2005, vol. XXXVI, pp. 117–122.
[46] I. Sargent, J. Harding, and M. Freeman, "Data quality in 3D: Gauging quality measures from users' requirements," in *Proc. Int. Symp. Spatial Quality*, Enschede, The Netherlands, 2007.

Konstantinos Karantzalos (M'05) received the engineering Diploma from the National Technical University of Athens (NTUA), Athens, Greece, in 2000 and the Ph.D. degree from NTUA in collaboration with Ecole Nationale de Ponts et Chaussees, Paris, France, in 2007. His thesis was entitled "Automatic feature extraction from aerial and satellite imagery with computer vision techniques."

In 2007, he was a Postdoctoral Researcher with the Medical Imaging and Computer Vision Group, Department of Applied Mathematics, Ecole Centrale de Paris, Chatenay-Malabry, France. He is currently a Lecturer of remote sensing with the Remote Sensing Laboratory, Department of Topography, Rural and Surveying Engineering, NTUA. His research interests include geoscience and remote sensing, computer vision, pattern recognition, artificial intelligence, and underwater photogrammetry. He has numerous publications in international journals and conferences.

Dr. Karantzalos is the recipient of the "Best Paper Award" in the International Symposium of Remote Sensing in 2006.

Nikos Paragios (SM'03) received the B.Sc (highest honors, valedictorian) and the M.Sc. degree (highest honors) in computer science from the University of Crete, Greece, in 1994 and 1996, respectively, the Ph.D. degree (highest honors) in electrical and computer engineering from Institut National de Recherche en Informatique et en Automatique (INRIA), France, in 2000, and the Habilitation a Diriger de Recherches - D.Sc. (HDR) degree from University of Nice, Sophia Antipolis, France, in 2005.

He was with Siemens Corporate Research, Princeton, NJ, in 1999–2004 as a Project Manager, Senior Research Scientist, and Research Scientist. In 2002, he was an Adjunct Professor with Rutgers University, Camden, NJ, and in 2004 with New York University. He was Professor/Research Scientist (2004–2005) with the Ecole Nationale de Ponts et Chaussees, Paris, France, He was a Visiting Professor with Yale University, New Haven, CT, in 2007. He is currently a Professor (Professeur des universites–premiere classe) with the Ecole Centrale de Paris, Chatenay-Malabry, France–one of most exclusive engineering schools "Grande Ecoles"–leading the Medical Imaging and Computer Vision Group, Applied Mathematics Department. He is also with the INRIA, Saclay Ile–de–France, Orsay, France, the French Research Institute in Informatics and Control, heading the GALEN group, a joint research team between ECP/INRIA. He has published more than 100 papers in the most prestigious journals and conferences of medical imaging and computer vision and has coedited four books. He is the holder of 15 U.S. issued patents with more than 20 pending.

Prof. Paragios is an Associate Editor for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Area Editor of the *Computer Vision and Image Understanding Journal* and member of the Editorial Board of the *International Journal of Computer Vision*, the *Medical Image Analysis Journal*, the *Journal of Mathematical Imaging and Vision*, and *Image and Vision Computing*. He is one of the program Chairs of the 11th European Conference in Computer Vision (ECCV'10, Heraklion, Crete). In 2008, he was the laureate of one of Greece's highest honor for young academics and scientists of nationality or descent (worldwide), and the recipient of the Bodossaki Foundation Prize in the field of applied sciences. In 2006, he was named one of the top 35 innovators in science and technology under the age of 35 from the MIT's *Technology Review* magazine.