Building Detection in a Single Remotely Sensed Image with a Point Process of Rectangles

Csaba Benedek, Xavier Descombes, Josiane Zerubia

Building Detection in a Single Remotely Sensed Image with a Point Process of Rectangles

Csaba Benedek[*][†], Xavier Descombes[*] and Josiane Zerubia[*]
[*]Ariana Project-Team INRIA/CNRS/UNSA, B.P. 93, 06902 Sophia Antipolis, France
[†]Distributed Events Analysis Research Group, Computer and Automation Research Institute
H-1111, Budapest, Kende utca 13-17, Hungary, bcsaba@sztaki.hu

Abstract

In this paper we introduce a probabilistic approach of building extraction in remotely sensed images. To cope with data heterogeneity we construct a flexible hierarchical framework which can create various building appearance models from different elementary feature based modules. A global optimization process attempts to find the optimal configuration of buildings, considering simultaneously the observed data, prior knowledge, and interactions between the neighboring building parts. The proposed method is evaluated on various aerial image sets containing more than 500 buildings, and the results are matched against two state-of-the-art techniques.

1 Introduction

Detecting buildings in aerial and satellite images [5, 6, 8] is a key issue in several remote sensing applications, among others in cartography, GIS data management and updating, disaster recovery or illegal built-up region detection. In lack of stereo based height information [6], building identification becomes a hard monocular object recognition task. Due to the quickly evolving spatial and spectral resolution of the images, the large variety of camera sensors, image quality, seasonal and weather circumstances, and the richness of the different building appearances it is extremely challenging to develop a widely applicable solution for the problem.

Most of the previous single view techniques are restricted to specific image properties and scene contents. They expect the fulfillment of various hypothesizes, such as buildings are homogenous areas either in color or in texture [7], roofs have unique colors which can distinguish them from the background [8], or shadows of buildings are present and can be extracted by color filtering [7, 8].

High contrast is often necessary to obtain a clear edge map for contour based detection [5, 8]. Other approaches assume that the building types in a given image set can be efficiently characterized by a couple of template buildings [4, 9], or one can apply simplified 3-D building structures composed of planar surfaces with parallel sides [5]. However combining the different solutions or adapting them to altered circumstances is not straightforward, although the recent remote sensing image databases demand to jointly handle highly heterogenous data. To ensure generality and robustness, besides extracting different limited descriptors, feature integration and selection should be addressed at the same time. Therefore we construct a method which can combine the features in a flexible way based on availability, enabling adaptation to various image sets.

In this paper we introduce a robust Marked Point process (MP) [3] model for the building detection problem. In Sec. 2, we describe the probabilistic framework of our approach, while Sec. 3 deals with feature modeling and integration. Evaluation and discussion are given in Sec. 4: the performance of the proposed model is compared to *two* reference methods through real aerial images containing 567, also manually validated, objects.

2 Marked Point Process Model

The input of the proposed framework is a single aerial or satellite image, which is modelled as a 2-D pixel lattice S, and $s \in S$ denotes a single pixel. $\mathcal{D}$ refers to the global image data. We assume that the footprint of each building can be approximated either as a rectangle or as the union of many slightly overlapping rectangular building segments, which we aim to extract by the following model. A building segment candidate u is described by five parameters: c_x and c_y center coordinates, e_L, e_l side lengths and $\theta \in [-90°, +90°]$ orientation [see Fig. 1(a)].

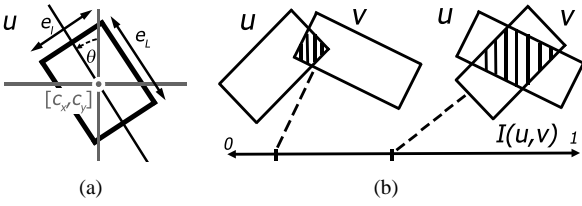Let be $\mathcal{H}$ the space of u objects. The Ω configuration

Figure 1. Demonstration of the (a) object rect-angle parameters and (b) calculation of the interaction potentials

space is defined as [3]:

$$\Omega = \bigcup_{n=0}^{\infty} \Omega_n, \quad \Omega_n = \left\{ \{u_1, \ldots, u_n\} \in \mathcal{H}^n \right\}$$

Denote by ω an arbitrary object configuration $\{u_1, \ldots, u_n\}$ in Ω. We define a $\sim$ neighborhood relation in $\mathcal{H}$: $u \sim v$ if their rectangles intersect.

We introduce a non-homogenous data-dependent Gibbs distribution on the configuration space: $P_{\mathcal{D}}(\omega) = 1/Z \cdot \exp[-\Phi_{\mathcal{D}}(\omega)]$, where $\Phi_{\mathcal{D}}(\omega)$ is called the configuration energy and Z is a normalizing constant. The energy is divided into data dependent ($A_{\mathcal{D}}$) and prior (I) parts:

$$\Phi_{\mathcal{D}}(\omega) = \sum_{u \in \omega} A_{\mathcal{D}}(u) + \gamma \cdot \sum_{\substack{u,v \in \omega \\ u \sim v}} I(u,v) \quad (1)$$

where $A_{\mathcal{D}}(u) \in [-1, 1]$, $I(u,v) \in [0, 1]$ and γ is a weighting factor between the two terms. The process searches for the maximum likelihood configuration estimate obtained as $\omega_{\mathrm{ML}} = \arg\min_{\omega \in \Omega} \left[\Phi_{\mathcal{D}}(\omega) \right]$.

The $A_{\mathcal{D}}(u)$ *unary* potential characterizes a proposed building segment $u = \{c_x, c_y, e_L, e_l, \theta\}$ depending on the local image data, but independently of other objects of the population. Rectangles with negative unary potentials are called *attractive objects*. Considering (1) we can observe that the optimal population should consist of attractive objects exclusively: if $A_{\mathcal{D}}(u) > 0$, removing u from the configuration results in a lower $\Phi_{\mathcal{D}}(\omega)$ global energy.

On the other hand, we have to avoid configurations which contain many objects in the same or strongly overlapping positions. Therefore, the $I(u,v)$ *interaction* potentials realize a prior geometrical constraint: they penalize intersection between different object rectangles [Fig. 1(b)]

$$I(u,v) = \frac{\#\{s | s \in u, s \in v\}}{\#\{s | s \in u\} + \#\{s | s \in v\}}$$

where $s \in u$ means that pixel s is covered by the rectangle of object u, and $\#$ refers to the cardinality of a set.

To fit the above framework to the building detection task, we need to handle two key issues. *Firstly*, an appropriate $\Phi_{\mathcal{D}}(\omega)$ energy function should be constructed where the ω_{ML} configuration efficiently estimates the true building population. Based on (1) this is primarily related with

the definition of the $A_{\mathcal{D}}(u)$ data term, thus we dedicate Sec. 3 to this problem. *Secondly*, we need to choose an optimization technique. We use the Multiple Birth and Death (MBD) algorithm [3] for this purpose, which evolves the population of buildings by alternating randomized object generation (*birth*) and removal (*death*) steps in a simulated annealing framework. Experimental evidences [3] show, that regarding computational complexity, MBD outperforms MCMC-based [6] relaxation algorithms, see details in [1, 2].

3 Flexible data term construction

This section deals with the construction of the $A_{\mathcal{D}}(u)$ data term. The process consists of three parts: feature extraction, energy calculation and feature integration. *First*, we define different $f : \{u, \mathcal{D}\} \to \mathbb{R}$ features which evaluate a building hypothesis for u in the image, so that 'high' f values correspond to efficient building candidates. We must consider here, that the decision based on a single f feature can lead to a *weak classification*, since the buildings and background may overlap in the f-domain. On the other hand, f might be an incomplete descriptor i.e. it can be relevant only for a group of buildings in the population.

In the test image of Fig. 3 three features are used. The `gradient` descriptor exploits that below the edges of a relevant rectangle candidate (R_u), we expect pixels (s) with large intensity gradient vectors (∇g_s) directing to the local normal vector ($\boldsymbol{n}_s$) of the rectangle. Therefore the Λ_u gradient descriptor is obtained as $\sum_{s \in \tilde{\partial} R_u} \nabla g_s \cdot \boldsymbol{n}_s$, where '$\cdot$' denotes scalar product and $\tilde{\partial} R_u$ is the dilated edge mask of rectangle R_u. The process is demonstrated in Fig. 3 (c)-(d).

The `shadow` feature is based on a preliminary cast shadow map (Fig. 3(e)). Exploiting that cast shadows are located next to the R_u object rectangles, one should check the presence of shadows in a parallelogram T_u^{sh} defined by R_u and the estimated sun direction vector, $\mathbf{d}$ [8] (Fig. 3(f)). The χ_u feature is calculated as the minimum of the filling ratio of shadowed pixels in T_u^{sh}, and the filling ratio of non-shadowed pixels in R_u.

Several roofs can be identified by their typical colors, for example pixels of *red tiles* have high a* color component values in CIE L*a*b* color space representation as shown in Fig. 3(g). Assume that based on a `roof color` hypothesis we can derive a binary mask image containing the estimated roof pixels e.g. by thresholding (Fig. 3(h)). Thereafter, we define the $\mathcal{C}_u$ color feature similarly to the shadow descriptor, prescribing high ratio of roof pixels inside R_u and low ratio in the region around R_u. Parameters can be set using Ground Truth data and conventional Maximum Likelihood estimation algorithms.

In the *second step*, we construct energy subterms for each $f \in \{\Lambda, \chi, \mathcal{C}\}$ feature, so that we attempt to satisfy $\varphi_f(u) < 0$ for real objects and $\varphi_f(u) > 0$ for false can-
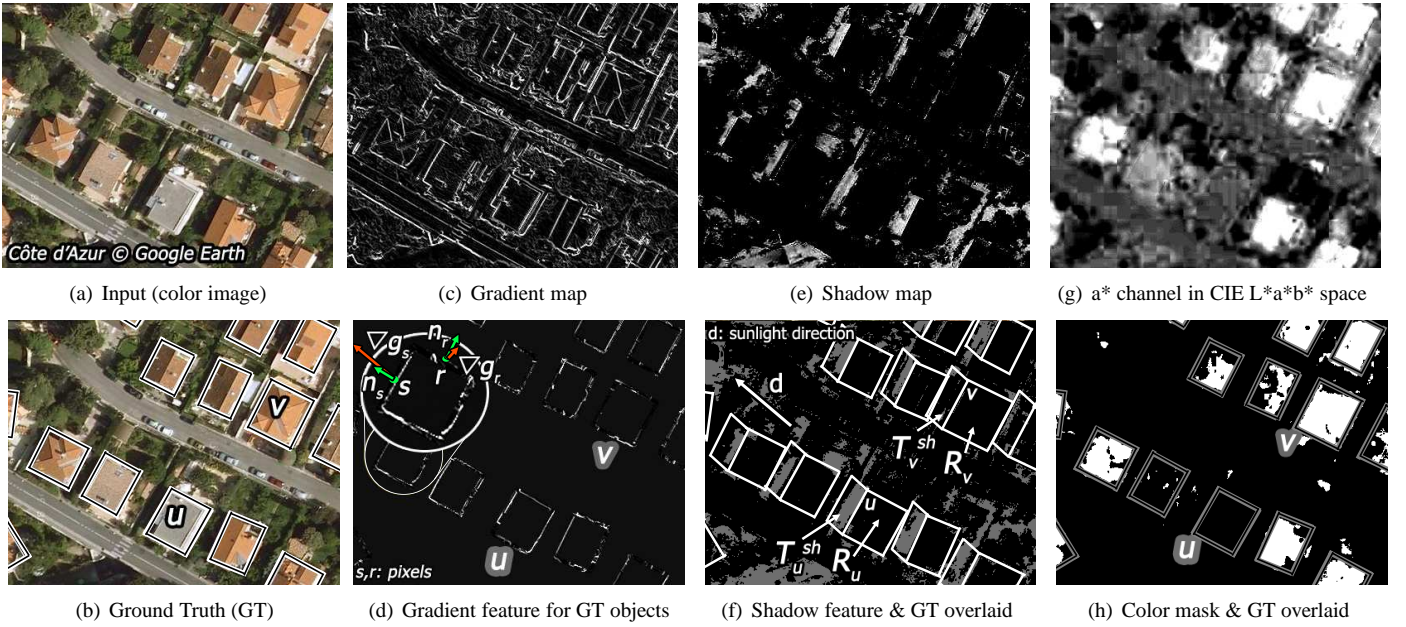
| (a) Input (color image) | (c) Gradient map | (e) Shadow map | (g) a* channel in CIE L*a*b* space |
| (b) Ground Truth (GT) | (d) Gradient feature for GT objects | (f) Shadow feature & GT overlaid | (h) Color mask & GT overlaid |

Figure 2. Feature maps of an image from the CÔTE D'AZUR **test set.**

didates. For this purpose, we project the feature domain to $[-1, 1]$ with a monotonously decreasing function:

$$\varphi_f(u) = \begin{cases} \left(1 - \frac{f(u)}{d_0^f}\right) & \text{if } f(u) < d_0^f \\ \exp\left(-\frac{f(u)-d_0^f}{D^f}\right) - 1 & \text{if } f(u) \geq d_0^f \end{cases}$$

where d_0^f and D^f are parameters. Consequently, object u is attractive according to the $\varphi_f(u)$ term iff $f(u) > d_0^f$, while D_f performs data-normalization.

Usually, the individual features are in themselves inappropriate to describe all buildings of the scene, which is illustrated in Fig. 2. We have chosen here two sample buildings segments u and v so that for u, the gradient and shadow features are efficient, while the roof color is irrelevant. The case of v is just the opposite. To handle such data heterogeneity, the proposed framework enables flexible *feature integration*. First, from the $\varphi_f(.)$ primitive terms introduced previously we construct different building prototypes. For each prototype we can prescribe the fulfillment of one or many feature constraints whose φ_f-subterms are connected with the max operator in the joint energy term of the prototypes (logical AND in the negative fitness domain). As well in a given image several building prototypes can be detected simultaneously if the prototype-energies are joined with the min (logical OR) operator. In our example, we use two prototypes: the first prescribes the edge and shadow constraints, the second one the roof color alone (as it is can detect the red roofs in itself accurately), thus the joint energy term is calculated as: $A_{\mathcal{D}}(u) = \min\left\{\max\left\{\varphi_{\Lambda}(u), \varphi_{\chi}(u)\right\}, \varphi_c(u)\right\}$.

4   Experiments

We evaluated our method on five aerial data sets obtained from Google Earth and the City Council of Budapest. To guarantee the heterogeneity of the test sets, we chose five completely different regions: Côte d'Azur (French Riviera), Normandy (FR), Manchester (UK), Bodensee (GER) and Budapest (HUN). We collected samples from densely populated suburban areas, and built a manually annotated database for the validation, containing 567 buildings.

For comparison, we have selected two methodologically different reference techniques from the literature: an Edge Verification (EV) method [8] and a Segment-Merge (SM) model [7]. We have focused on validating the model structures instead of special input-dependent descriptors, thus we have taken care of choosing references which use similar image features (gradient, shadow, color) to our framework, but they exploit them in different manners. More precisely, in EV [8], the shadow and roof color information is only used to coarsely detect the built-in areas, while the object verification is purely based on matching the edges of the building candidates to the Canny edge map extracted over the estimated built-in regions. On the other hand, the SM model iterates three steps: (i) building segment estimation by seeded region growing, (ii) region merging and shadow evidence verification, and (iii) filtering based on geometric and photometric features.

For a sample image, Fig. 3 shows detection results with the three methods (EV, SM and the proposed MP) and the Ground Truth (GT) configuration. In the quantitative evaluation we counted the number of missing and falsely detected
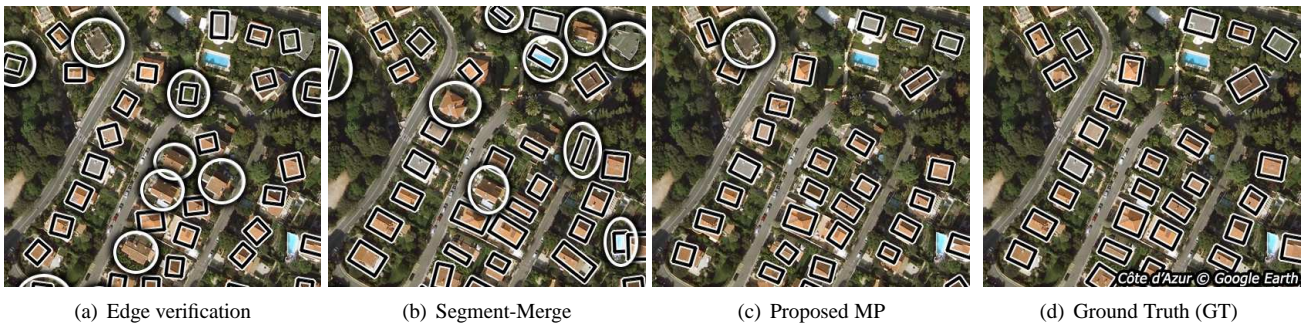
(a) Edge verification   (b) Segment-Merge   (c) Proposed MP   (d) Ground Truth (GT)

Figure 3. Evaluation (from the CÔTE D'AZUR set): comparing the MP model to the EV technique [8] and the SM method [7]. Circles denote completely missing or false objects.

objects, results are provided in Table 1 (in the last row, the error rates are given in percent of the population).

We continue with the discussion. Since both the EV and SM reference methods follow the deterministic object generation-acceptance scheme, buildings ignored in the hypothesis generator phase appear automatically as missing objects (see Fig 3 (a) and (b)). On the contrary, the introduced MP model proposes buildings in a stochastic way, thus objects *can be* generated with any position and appearance parameters. The acceptance depends on the robust inverse object description in the energy model, while the computational tractability is ensured by optimized relaxation parameters [3] and a non-uniform birth process [2].

Another important observation is that the EV and SM methods are sequential, thus the failure of each step may cause a bottleneck for the whole process, e.g. due to a weak edge map, missing shadows or overlapping color domains. On the contrary, the proposed model uses different prototype-hypothesizes parallely, thus they may enable to detect the buildings even in cases of partially missing or irrelevant feature information. Results in Fig. 3 and Table 1 confirm the generality of the proposed model and its superiority versus the EV and SM approaches.

5 Conclusion

We have proposed a Marked Point Process framework for building extraction in a single remotely sensed image. The method implements a flexible hierarchical feature integration scheme to characterize different buildings based on different feature-tuples. The evaluation confirmed the advantages of the approach in various building datasets.

References

[1] C. Benedek, X. Descombes, and J. Zerubia. Building extraction and change detection in multitemporal aerial and satellite images in a joint stochastic approach. Research Report 7143, INRIA, Sophia Antipolis, December 2009.

Table 1. Numerical comparison of the EV [8], the SM [7] and the proposed methods (MP).

Data Set	#B*	Missing objects			False objects		
		EV	SM	**MP**	EV	SM	**MP**
CÔTE D'AZUR	123	14	20	5	20	25	4
BODENSEE	80	11	18	7	13	15	6
BUDAPEST	41	11	9	2	5	1	4
NORMANDY	152	18	30	18	32	58	1
MANCHESTER	171	46	53	19	17	42	6
ALL (%**)	567	18%	23%	9%	15%	25%	4%

*#B denotes the number of buildings in the test sets
**the missing/false objects are given in percent of #B

[2] C. Benedek, X. Descombes, and J. Zerubia. Building extraction and change detection in multitemporal remotely sensed images with multiple birth and death dynamics. In *IEEE WACV*, pages 100–105, Snowbird, USA, 2009.

[3] X. Descombes, R. Minlos, and E. Zhizhina. Object extraction using a stochastic birth-and-death dynamics in continuum. *J. Mathematical Imaging and Vision*, 33:347–359, 2009.

[4] K. Karantzalos and N. Paragios. Recognition-driven two-dimensional competing priors toward automatic and accurate building detection. *IEEE Trans. GRS*, 47(1):133–144, 2009.

[5] A. Katartzis and H. Sahli. A stochastic framework for the identification of building rooftops using a single remote sensing image. *IEEE Trans. GRS*, 46(1):259–271, 2008.

[6] F. Lafarge, X. Descombes, J. Zerubia, and M. Pierrot-Deseilligny. Structural approach for building reconstruction from a single DSM. *IEEE Trans. PAMI*, 32(1):135–147, 2009.

[7] S. Muller and D. Zaum. Robust building detection in aerial images. In *CMRT*, pages 143–148, Vienna, Austria, 2005.

[8] B. Sirmacek and C. Unsalan. Building detection from aerial imagery using invariant color features and shadow information. In *ISCIS*, Istanbul, Turkey, 2008. [CD-ROM].

[9] B. Sirmacek and C. Unsalan. Urban-area and building detection using SIFT keypoints and graph theory. *IEEE Trans. GRS*, 47(4):1156–1167, April 2009.