COMBINING TOP-DOWN AND BOTTOM-UP APPROACHES FOR BUILDING DETECTION IN A SINGLE VERY HIGH RESOLUTION SATELLITE IMAGE

Mahmoud Mohammed Sidi Youssef [1,2], *Clément Mallet* [1], *Nesrine Chehata* [3,4], *Arnaud Le Bris* [1], *Adrien Gressin* [1]

1. IGN/MATIS lab., Université Paris Est, France
2. Sup'Com Ecole Supérieure des Communications de Tunis, Tunisia
3. Laboratoire G&E (EA 4592), IPB / Université de Bordeaux, France
4. IRD/UMR LISAH El Menzah 4, Tunis, Tunisia

ABSTRACT

Building detection from geospatial optical images has been a popular topic of research for the last twenty years and in particular with the emergence of very high resolution satellites. Existing methods exhibit various flaws and prevent them from being efficient at large scales of space and time: they are context-dependent, require a tedious parameter tuning or several data sources. In this paper, we propose a fully automatic method that alleviates some of these issues by combining the strengths of bottom-up and top-down approaches, *i.e.*, of both classification and pattern recognition algorithms. This allows to correctly detect the objects by geometric prior knowledge while finely delineating their borders and preserving their shapes. The method is evaluated over a complex area of more than 230 buildings using a 0.5 m multispectral pansharpened Pleiades image.

***Index Terms*—** Segmentation, classification, Marked Point Process, building, very high resolution imagery.

1. INTRODUCTION

Building detection in very high resolution (VHR) satellite images is a key issue for numerous remote sensing applications. The identification of buildings has interested the scientific community for the two last decades using different kinds of data such as multispectral or hyperspectral optical images, SAR data, LIDAR data, sometimes eased by 3D information using optical multiple-view images and external data such as cadastral maps [1]. Here, this paper is devoted to building detection from a single VHR optical satellite image, which offers a suitable trade-off between spatial, spectral and temporal resolutions.

Most of the previous single-view techniques are restricted to specific image properties and scene contents. They expect the fulfilment of various hypothesis, such as "buildings are homogeneous areas either in color or in texture", "roofs have unique colors which can distinguish them from the background", or "building shadows are present and can be extracted by color filtering".

Existing works can be grouped into three categories. Bottom-Up (BU) approaches are based on low-level feature computation. The spectral and spatial contents of submetric optical images provide a detailed description of urban scenes, which is perfect for fine building detection [2]. However, such content may lead to poor or noisy results due to the high scene heterogeneity and complexity. Conversely, Top-Down (TD) methods directly focus on object detection, using prior knowledge on building shapes, spatial arrangements and interactions between objects [3, 4]. However, proposed methods are often complex, leading to multiple parameters and significant computing times for large areas. Finally, mixed methods aim to benefit from advantages of both approaches. Generally, graph-based approaches are selected so as to propagate local feature-based classifications to large areas with structured reasoning [5]. They are efficient, but finding the optimal building configuration requires a global optimization framework, redhibitory over large scales. In addition, they remain most of the time context-dependent. Consequently, another solution is to alternatively perform BU and TD detections [6].

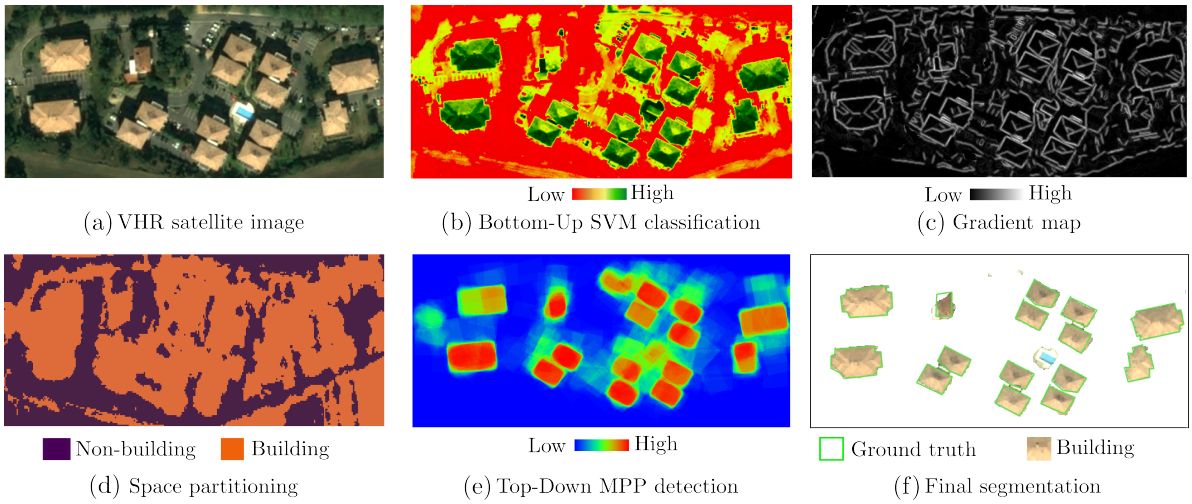
In this paper, we propose such kind of approach that alleviates some key issues mentioned above by simply and efficiently combining BU and TD approaches.

2. METHODOLOGY

The proposed approach for building detection is illustrated in Figure 1, and can be decomposed into three main steps.

1. A **Bottom-Up** (BU) approach which exploits the image information at the pixel level. It allows to derive two kinds of cues for building detection. On the one hand, a per-pixel supervised classification is performed and a probability map of buildings is computed. On the other hand, low-level primitives are extracted so as to support, improve and accelerate the Top-Dow detection of rectangular patterns.

2. A **Top-Down** (TD) extraction step that also aims to provide a probability map of buildings. Such map is retrieved as the



(a) VHR satellite image (b) Bottom-Up SVM classification (c) Gradient map

(d) Space partitioning (e) Top-Down MPP detection (f) Final segmentation

Fig. 1. Overview and sample result of our approach.

merge of multiple extractions of rectangular patterns using Marked Point Processes (MPP). Since building may have various sizes and shapes, and knowing that rectangles only allow to find the coarse location of buildings, several hypothesis are formulated and allow to better delineate building regions.

3. A **fusion/decision step** that merges BU classification with TD detection. A pixel-based binary map is computed using a Markov Random Field (MRF) formulation, optimized using the graph-cuts framework. It benefits from advantages of both previous steps: a strong but coarse evidence of buildings (TD approach) coupled with a fine but noisy location of their edges (BU approach).

Our approach is similar to the work of [6] but with several differences. In [6], BU (Markov Random Field unsupervised classification) and TD (MPP rectangle extraction) are alternatively performed until a convergent classification is obtained. Each step alternatively takes advantage of the other one. In practice, only few pixels are swapped after few iterations. In our paper, we consider that the BU step is sufficiently reliable to perform a single refinement step. Furthermore, the succession of optimization steps between both levels is substituted by merging multiple TD binary rectangle maps which is the genuine novelty in this paper.

2.1. Bottom-Up scene analysis

The first step consists in performing a per-pixel classification of the satellite image. Two classes are considered: *building* and *non-building*, which is sufficient for correct discrimination. No spatial smoothing is required since such regularization will be addressed by the final merging approach with the integration of the Top-Down knowledge. A supervised Support Vector Machines classification is adopted to better capture the various appearances of building roofs in images. Only the four spectral bands of the satellite image are used as features (*red*, *green*, *blue*, and *infra-red* channels), and are sufficient to focus on the most probable candidate regions for building extraction. A probability map of buildings is generated using the confidence values provided by the SVM (Figure 1b).

Secondly, low-level cues are extracted so as to ease the Top-Down extraction of rectangles. It is supported by color gradients, detected in the images. For that purpose, an enhanced gradient map is processed which gives the strength of the gradient at each pixel location. It is obtained by combining a basic gradient map with an unexhaustive but robust and reliable line segment detector, namely LSD [7]. This approach may lead to numerous false positives lying inside building roofs, vegetation and shadow contours as well as other highly contrasted ground items (see Figure 1c). However, they will be pruned by subsequent rectangle extraction. Furthermore, the detection of rectangles is based on a stochastic approach which extensively explores the space of associations of rectangles so as to find the configuration that best fits to the low-level cues extracted from the images. The process can be accelerated by limiting the search space to the most relevant areas of the images using a data-driven approach: the probability map of buildings is smoothed and binarized using object-based analysis. The satellite image is segmented and regions with high NDVI values (likely to correspond to vegetation) are no longer considered for the TD step (Figure 1d). Even if such discrimination is not perfect, up to 40% of an area of interest can be discarded.

2.2. Top-Down building extraction

It is assumed that each building footprint can be approximated either as a rectangle or as the union of several slightly overlapping rectangles. The coarse localization of the buildings is obtained by matching this simple model to the image and subsequently provides rough regions where the foreground (*i.e.*

the objects) and background are present. A stochastic framework is used to model the image with a set of rectangles. Each configuration of such patterns is measured by an energy. This energy computes the consistency between the configuration and the observed image (namely the enhanced gradient map where vegetation areas have been discarded), and takes into account interactions between neighbouring rectangles (they should not overlap too much). The energy minimization is complex since the number of objects in the configuration is unknown, and the energy is not convex. Most of conventional optimization algorithms cannot be performed in such conditions. The Marked Point Process framework perfectly suits to that purpose and has been adopted [8]. Point processes are usually simulated using a Reversible Jump Monte Carlo Markov Chain (RJMCMC) sampler coupled with a stochastic relaxation [9].

Our model is efficient in retrieving main building parts but exhibit two limitations. On the one hand, erroneous (*e.g.,* park lots, shadows) and imprecise detections may happen. On the other hand, results are sensitive to parameter tuning. This prevents the direct use of the output of a single optimization as TD probability map for final classification. Consequently, a large range of parameters is tested and their results are accumulated (in our experiments, 24 distinct results, see Figure 1e). Finally, each pixel of the image is associated to a probability based on the number of rectangles it belongs to. One can see that rectangle stacking allows to provide a reliable foreground/background map. Low confidence values correspond to misdetection and building edges whereas highest values correspond to building roofs.

2.3. BU and TD combination for final segmentation

For the final binary decision process, a non supervised approach is adopted. An energy minimization is proposed to classify the pixels. The energy is the sum of a data component, derived from the Bottom-Up classification process, and a prior term, describing pairwise interaction. The latter one is defined from the building probability measure of the Top-Down extraction step. It aims to favor neighboring pixels with similar foreground probabilities, and strongly penalizes pairs with highly contrasted values. Therefore, it acts as a smoothing term that respects TD breaking lines and rectangular shapes.

The proposed energy is graph-representable. A Graph-Cut based algorithm [10] is reach the global optimum of the energy (our model fits the requirements for this algorithm). Therefore, the proposed method is both a fusion and a regularization method (see [11] for more details).

3. RESULTS AND DISCUSSION

The method was evaluated on a mixed urban - rural area of the city of Toulouse (France). It covers almost 1.5 km^2, with 230 buildings over a hilly terrain. Vegetation is also present and consists of both woods and fields. The four-channel (red, green, blue, and infra-red) image is a simulation of Pleiades satellite data, with a spatial resolution of 0.5 m in the panchromatic mode and 2 m in the multispectral one. The area is rather challenging since buildings exhibit a heterogeneous behaviour in terms of size and appearance. In particular, various roof materials exist and grey ones (concrete and slant roofs) can be easily confused with road regions. For detailed analysis, 3 areas of interest were selected.

Area	Completeness (%)	Correctness (%)
1	89	81.5
2	90.5	80.5
3	94	67

Table 1. Quality assessment of our building extraction pipeline, compared to an existing 2D topographic database.

The computing time for each area is around 50 minutes and is mainly due to the MPP-based multiple rectangle extraction (BU classification and final graph-cut optimization being almost instantaneous). Each MPP corresponds to 15 million iterations and takes approximatively 2 minutes [8].

Figures 1 and 2 show results for the three areas of interest. Since a basic SVM classification with only four spectral channels was used, one can see that the BU classification is very sensitive to shadowed rooftops (leading to different BU probabilites per rooftop) and to roof materials. In addition some misclassification may occur between grey rooftops and roads as they have similar radiometry. One can also see that TD probability map is robust to shadows and highlights buildings with strong evidence even if only coarse locations are retrieved. Most of the buildings are detected, even those lying in forested areas. However, some false positives still remain. They correspond to ground areas with high linear contrast such as swimming pools and vehicles queued in park lots. In both cases, the BU approach helped minimizing these errors and more advanced models in BU and TD approaches should solve these issues.

The proposed approach accurately delineates most of the buildings. A quantitative analysis has been performed to assess the quality of the proposed approach. Reference footprints have been extracted from an imperfect existing 2D topographic database. The resulting *building/non-building* binary image is compared to the output of our automatic process. Completeness and correctness were calculated on a per-pixel evaluation basis and are provided in Table 1.

4. CONCLUSION AND PERSPECTIVES

A simple yet efficient workflow for combining standard Top-Down and Bottom-Up approaches was proposed for the well-known problem of building detection from monoscopic VHR

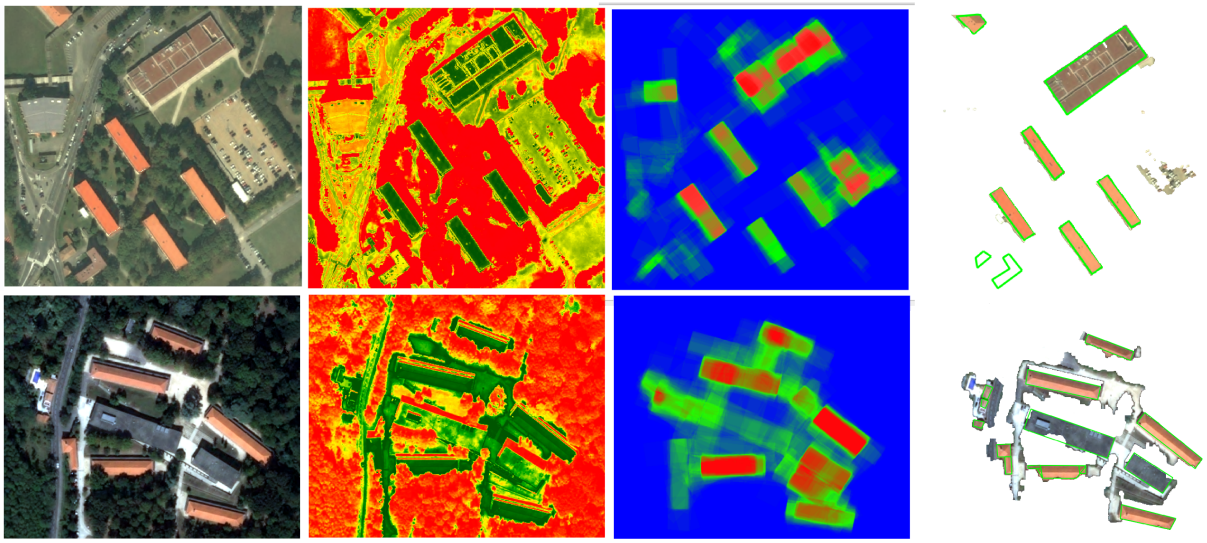


Fig. 2. Results for 2 areas of interest. **Top and Bottom**: *Area 2* and *Area 3*. **From left to right**: Satelliet image – BU building probability map – TD MPP building probability map – Final segmentation (see Figure 1 for the color codes).

satellite images. The novelty came from the adoption of state-of-the-art techniques for solving each step of the workflow, and the proposal of a solution with few prior knowledge and parameters. In addition, through the accumulation of MPP results, we managed to alleviate the standard issue of parameter tuning of this stochastic approach. Finally, it was based on open-source softwares [8, 12] that makes the work easily reproducible. Results are satisfactory both in terms of building delineation and detection. More images will be processed in order to assess the versatility of the approach, both in terms of landscapes and sensors. Finally, our appproach is currently benchmarked with other state-of-the-art methods [1, 6, 13], and results will be reported in forthcoming papers.

5. REFERENCES

[1] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS Journal*, vol. 86, pp. 21–40, 2013.

[2] C. Senaras, M. Ozay, and F.T. Yarman Vural, "Building detection with decision fusion," *IEEE JSTARS*, vol. 6, no. 3, pp. 1295–1304, 2013.

[3] C. Benedek, X. Descombes, and J. Zerubia, "Building detection in a single remotely sensed image with a point process of rectangles," in *ICPR*, 2010, pp. 1417–1420.

[4] K. Karantzalos and N. Paragios, "Recognition-driven two-dimensional competing priors toward automatic and accurate building detection," *IEEE TGRS*, vol. 47, no. 1, pp. 133–144, 2009.

[5] B. Sirmacek and C. Unsalan, "Urban area and building detection using SIFT keypoints and graph theory," *IEEE TGRS*, vol. 47, no. 4, pp. 1156–1167, 2009.

[6] D. Chai, W. Förstner, and M. Ying Yang, "Combine Markov Random Fields and Marked Point Processes to extract building from remotely sensed images," *ISPRS Annals*, vol. I-3, pp. 365–370, 2012.

[7] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," *IEEE TPAMI*, vol. 32, no. 4, pp. 722–732, 2010.

[8] M. Brédif and O. Tournaire, "LibRJMCMC: An open-source generic C++ library for stochastic optimization," *ISPRS Archives*, vol. XXXIX-B3, pp. 259–264, 2012.

[9] X. Descombes, Ed., *Stochastic Geometry for Image Analysis*, Wiley, 2011.

[10] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," in *ICCV*, 1999, pp. 377–384.

[11] M.M. Sidi Youssef, C. Mallet, N. Chehata, A. Le Bris, and A. Gressin, "Détection de bâtiments à partir d'une image satellitaire par combinaison dapproches ascendante et descendante," in *RFIA*, 2014.

[12] J. Inglada and E. Christophe, "The Orfeo Toolbox remote sensing image processing software," in *IGARSS*, 2009, pp. 733–736.

[13] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Contextual classification of lidar data and building object detection in urban areas," *ISPRS Journal*, vol. 87, pp. 152–165, 2013.