Robust Rooftop Extraction From Visible Band Images Using Higher Order CRF

Er Li, *Member, IEEE*, John Femiani, *Member, IEEE*, Shibiao Xu, *Member, IEEE*, Xiaopeng Zhang, *Member, IEEE*, and Peter Wonka, *Member, IEEE*

***Abstract*—In this paper, we propose a robust framework for building extraction in visible band images. We first get an initial classification of the pixels based on an unsupervised presegmentation. Then, we develop a novel conditional random field (CRF) formulation to achieve accurate rooftops extraction, which incorporates pixel-level information and segment-level information for the identification of rooftops. Comparing with the commonly used CRF model, a higher order potential defined on segment is added in our model, by exploiting region consistency and shape feature at segment level. Our experiments show that the proposed higher order CRF model outperforms the state-of-the-art methods both at pixel and object levels on rooftops with complex structures and sizes in challenging environments.**

***Index Terms*—Buildings, rooftops conditional random field (CRF), shadows.**

I. Introduction

EXTRACTED rooftops from remote sensing images play a prominent role in widespread applications, such as urban planning, 3-D city modeling, and flight simulation. While enormous advances have been made on building detection over the last years [1], it remains a challenging task to develop generic and robust algorithms. This is because the appearance of rooftops varies due to many factors, e.g., lighting conditions, a variety of reflections, diversity of image resolution, and image quality.

Most existing approaches identify rooftops by exploring image features based on several simplifying assumptions. As man-made objects, rooftops are often decomposed into simple geometric parts, with uniform color distribution within a single rooftop, and high contrast with surroundings [2]. While the effectiveness of all these properties has been demonstrated in prior work [3], the problem lies in the uncertainty of both the features and assumptions. On one hand, the assumptions are not always true for each rooftop. For example, many methods assume that rooftops have rectangle shape (see [4]–[6]), wherein conflicts with complicated building structures may be observed in real imagery.

David Marr's theory of vision [7] viewed typical visual recognition processing as a bottom-up hierarchy in which information is processed sequentially with increasing complexity. When this approach is applied to rooftops, lower level information tells us where objects are, and higher level information tells us which objects form a rooftop. Superpixels were introduced by Ren and Malik [8] in order to cluster pixels into atomic regions with homogeneous size and shape. Our work is based on the idea that these regions need not be atomic, but they are still useful in guiding image segmentation.

Based on those observations, we propose a novel rooftop extraction method using a higher order conditional random field (HCRF). The basic idea of our approach is to combine the high-level information (obtained from segments by presegmenting the aerial image) and the low-level information (pixels in the image) by using HCRF during the extraction. Existing methods, which only utilize segments or pixels, have several weaknesses: 1) segment-based methods tend to be highly sensitive to the accuracy of the initial segments, which is also a challenging question; 2) pixel-based methods fail to capture the global structure information across the whole image. For example, roads usually can be ruled out due to their long and thin shape; however, pixel-based methods do not have any knowledge about this, and thus, sometimes parts of road are mislabeled as rooftops. Another problem of pixel-based methods is that the commonly used standard conditional random field (CRF) models (see details in Section III-A4) often produce overly smooth segmentation results, which will merge closely spaced rooftops together. Moreover, existing methods (see [9]–[11]) suffer from incorrect shadows and vegetation detection before rooftops extraction, particularly when only RGB information is available. We demonstrate how to use segments to extract shadows and vegetation robustly.

We demonstrate the effectiveness of our approach on the SZTAKI-INRIA benchmark [12] and show that our higher order model improves pixel-level, as well as object-level, accuracy without any training data. Our contributions include the following.

- We propose a novel HCRF-based method, which incorporates both pixel-and segment-level information for the segmentation of rooftops. High accuracy is achieved by exploiting color features at the pixel level, along with region consistency and shape features at the segment level.

Manuscript received June 27, 2014; revised October 25, 2014 and December 24, 2014; accepted January 11, 2015. This work was supported by the National Natural Science Foundation of China under Grant 61331018, Grant 91338202, and Grant 61100132.

E. Li and J. Femiani are with the Department of Engineering and Computing Systems, Arizona State University, Mesa, AZ 85212 USA (e-mail: erli2@asu.edu; john.femiani@asu.edu).

S. Xu and X. Zhang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: shibiao.xu@nlpr.ia.ac.cn; xpzhang@nlpr.ia.ac.cn).

P. Wonka is with the Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 852871 USA, and also with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia.

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.2015.2400462

- We present a simple but robust shadows and vegetation extraction method based on the proposed framework.

II. Prior Work

A significant amount of work has been done on extracting buildings from aerial and satellite images due to the widespread use of rooftop footprints in geospatial applications. We refer the reader to [1] for a detailed discussion. Here, we only review the most relevant work.

One popular way of extracting buildings is exploring their shapes. It is observed that rooftops have more regular shapes, which usually are rectangular or combinations of several rectangles. Based on this, Liu *et al.* [4] applied multiresolution segmentation on fused images, extracting rectangular building roofs by reconstruction and multiscale classification from the roof polygon primitives. Cui *et al.* [5] used Hough transformation to extract perpendicular and parallel lines, which comprise the structure of a building. Then, region information is incorporated to construct a building structural graph, and the boundary of buildings is finally extracted from a cycle detection on the graph. Liu *et al.* [6] proposed a general semiautomatic rectilinear rooftop extraction method based on localized multiscale object-oriented segmentation and model matching. Shape-based approaches heavily rely on the accuracy and completeness of the extracted contour of rooftops. This, however, is also a difficult problem, and some of the methods [5], [6] assume that the rooftops in one image have the same orientation, which is not always true either. Müller and Zaum [13] found homogeneous roof candidate regions by seeded region growing. This initial segmentation then becomes the basis of several following steps. However, building extraction based solely on an initial segmentation tends to suffer from the accuracy of the initial segmentation.

Instead of classifying the shapes of rooftops directly, some researchers try to detect the rooftops based on their strong corners and edges. Martinez Fonte *et al.* [14] revealed that corner detectors could provide distinctive information on the type of structure in a satellite image, but they did not provide a complete process to extract rooftops from satellite image using corners. Sirmacek and Ünsalan [15] utilized scale invariant feature transform (SIFT) and graph theoretical tools to extract buildings from urban area. However, their method needs specific building templates for the subgraph matching. Then, in [16], they further proposed to build their extraction based on local feature vectors, which might lead to false positives when redundant local features appear in the image. Katartzis and Sahli [17] constructed hypothesis graph based on the edges of buildings and then used a Markov random field model to describe the dependencies between available hypotheses with regard to a globally consistent interpretation. However, the detection of edges is sensitive to the resolution and noise of the image. Nosrati and Saeedi [18] proposed to use edge definitions and their relationships with each other to create a set of potential vertices. Then, polygonal rooftops are extracted by studying the relationship between these potential vertices. The method lacks the ability of capturing nonpolygonal curved rooftops. Cote and Saeedi [11] further developed an automatic rooftop detection method in nadir color aerial imagery by combining corners and variational level set evolution method. The corners are assessed using multiple color-invariant spaces, and then, rooftop outlines are produced from selected corners through level set curve evolution. Their method cannot distinguish rooftops from other structures with salient boundaries in the image, and the selection of parameters tends to lack robustness under varying resolutions.

Shadows are another significant feature of buildings. Several authors use shadows for hypothesis verification and height estimation [19]–[23] after an initial building detection step. Liow and Pavlidis [24] used shadows for hypothesis verification. They first extracted line segments, and regions that lie next to shadows were considered for building hypothesis. To get the final rooftop, a region growing algorithm is used to find other edges of the rooftops. However, this edge-based method often suffers from the ambiguity between edges produced by strong ridges of gable and hip roofs and actual edges of building. Sirmacek and Ünsalan [25] used invariant color features to extract information from shadows and then determine the illumination direction and verify building location based on shadows. Finally, a rectangle fitting method is used to align a rectangle with the canny edges of the image. This method is sensitive to the edge quality and limited to rectangular buildings. Akcay and Aksoy [9] first proposed using shadows and directional spatial constraints to detect candidate building regions. However, the building regions are selected by clustering the candidate patches from an initial oversegmentation, which might not be correct in each patch. Recently, Ok *et al.* [10] have adopted the idea of using shadows and directional spatial constraints and proposed to extract the final rooftop using CRF optimization at pixel level. They dilated the shadows along the opposite of light direction in a certain distance to obtain a region of interest (ROI) for each rooftop. Then, they ran CRF in each ROI to label pixels inside it as rooftops or nonrooftops. This ROI-based method can break one rooftop into separated pieces when the shadows of buildings are incomplete due to clutter or vegetation near the buildings. To overcome this problem, in [1], they further proposed to run a multilabel CRF segmentation over the whole image after getting an initial segmentation following the method in [10]. Since the initial segmentation is estimated based only on shadows, the accuracy of method in [10] notably decreases when extracting shadows is not reliable. Femiani and Li [26] and Femiani *et al.* [27] extended this graph-based approach by showing how additional sources of geospatial data can be used to guide the shadow-based segmentation, as well as by verifying that extracted features cast shadows and resegmenting the image until a set of rooftops are extracted that are consistent with visible shadows. Different from the CRF at pixel level, Wegner *et al.* [28] tried to run CRF at higher level by constructing the graph of CRF on image segments instead of image pixels. Benedek *et al.* [12] integrated different local features into object-level features and then adopted Markov random field at object level. These two methods run stochastic object extraction merely on higher level, which impairs accuracy at pixel level. By contrast, in this paper, we proposed a novel HCRF model, which combines both pixel- and segment-level information during the stochastic object extraction process. Comparing with pixel-level-based

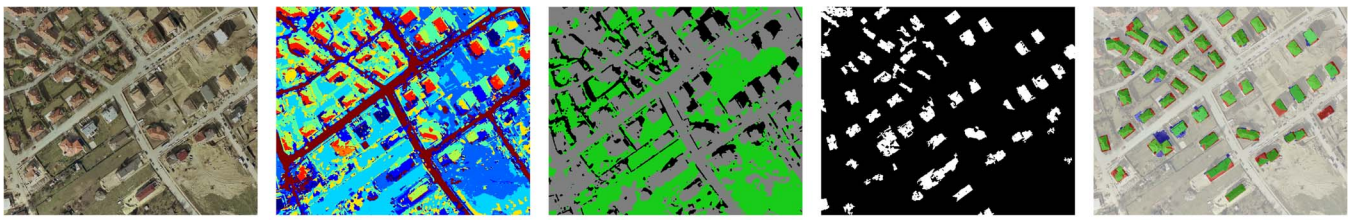

Fig. 1. Overview of the proposed method. (a) Starting from an aerial image, we (b) decompose the input data into segments. Then, (c) vegetation (green) and shadows (black) are extracted, and (d) probable rooftops are selected from the segments. Finally, (e) an accurate segmentation for the rooftops is provided by using an HCRF optimization process. Correct results are shown in green, false positives are shown in blue, and false negatives are shown in red.

methods [1], [10], [27], a new higher order potential defined at the segment level is added in our model, which enables it to encode high-level structure, and thus improves the robustness of extraction, and unlike those methods [12], [28] that merely run stochastic object extraction at segment level, pixel-level information is also taken into account in our model, which improves the accuracy of extraction.

III. Segmentation Algorithm

The proposed method takes as input a remote sensing image with only RGB information. The goal is to extract rooftops from a single image as accurately as possible without any user interaction. The main idea of our approach is to integrate pixel- and segment-level information for extraction of rooftops through the proposed novel HCRF model. The whole method consists of the following key steps.

1) First, the given image will be segmented into several patches by unsupervised clustering [see Fig. 1(b)].
2) Based on the segments from the first step, we will extract vegetation and shadows [see Fig. 1(c)].
3) Remaining unlabeled patches will be classified into probable rooftops and probable nonrooftops depending on shape, size, compactness, and shadows [see Fig. 1(d)].
4) A higher order multilabel CRF segmentation is performed to get final results [see Fig. 1(e)].

A. Algorithm

1) Initial Presegmentation: To obtain high-level information, we first decompose the image into basic elements that preserve the relevant structure of the objects in the image. We aim to cluster pixels into perceptually homogeneous regions, which should ideally correspond to different real-world objects in the aerial image (e.g., trees, roads, and rooftops). To achieve this type of decomposition, we use Gaussian mixture model (GMM) clustering method to segment the image into homogeneous regions. GMM assumes the underlying data to belong to a mixture of Gaussian distributions. Specifically, the probability density function is expressed as a weighted sum of M component Gaussian densities, i.e.,

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} \omega_i g(\mathbf{x}|\mu_i, \sigma_i) \tag{1}$$

where ω_i represents the mixture weight; and μ_i, σ_i denote the mean vector and the covariance matrix for each component,

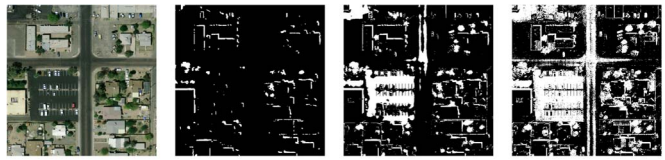

Fig. 2. Comparisons between the proposed shadow extraction method and two other typical methods. (a) Original image. (b) Our result. (c) Result of [33]. (d) Result of [34]. White areas represent the detected shadows.

respectively. Despite its simplicity, GMMs with a relatively small number of mixture components have proven to be excellent performer in modeling natural image patches [29]. During our experiment, we found that $M = 10$ works well on most of our data sets. We use a full-covariance GMM for the initial segmentation, following the guidance in [30]. In practice, we first convert the RGB channel of an aerial image into Lab color space [31]. A median filter is used to reduce the noise in the image before the segmentation, and then, the expectation-maximization algorithm [32] is used to fit the mixture model.

2) Extract Shadows and Vegetation From GMM Labels: Compared with rooftops and ground, shadows and green vegetation have significant features that can be identified in the aerial imagery more easily. Extraction of shadows and vegetation can help to rule out those regions that are not likely to be rooftops. Shadows are also a strong clue of the existence of buildings nearby, which can be used to localize probable rooftops. Although a lot of work has been done on extracting shadows (see [33] and [34]) and vegetation (see [35] and [36]) from remote sensing images, the methods of [1] and [10] require near infrared (NIR) to achieve high performance, and the accuracy decreases significantly if there is only RGB channel. One failure case of two typical shadows extraction methods is shown in Fig. 2. It is hard for an automatic approach to select an optimal threshold only at pixel level due to the noise and other dark regions such as roads.

Based on the preceding observations, we propose a robust and automatic shadows extraction method by utilizing the presegmentation result from Section III-A1. Given an image I and label class $L = 1, 2, \ldots, 10$, each pixel is assigned a label L_i from L in Section III-A1, and all of the pixels are classified into ten classes. We select the mean intensity of the class with the lowest mean intensity as the initial threshold to identify shadows. For vegetation, we use the color index proposed in [35] instead of intensity. This initial threshold is

Fig. 3. How to remove nonprobable rooftops step by step. (a) Original image. (b) Shadows and vegetation are removed. (c) Remaining blobs are further removed depending on their size, eccentricity, and compactness. (d) Final probable rooftops after pruning out the blobs in (c) using shadows and light direction information. (e) Ground truth of rooftops.

sufficient to extract most of the shadows and vegetation, but it is not sufficient to recover all of them using a global threshold due to noise and variation of the appearance of shadows and vegetation. To overcome this problem, we further improve the shadows extraction result by performing multilabel graph cut [1], starting from the initial shadow and vegetation mask.

3) Extracting Probable Rooftops: Usually, probable rooftops can be localized from shadows and light direction following the methods in [10]. The basic idea is generating ROI by shifting the shadows in the opposite of light direction through a certain distance, and then, the probability of each pixel in that ROI belonging to a rooftop is determined according to some rules. The efficiency of this shadow-based method has experimental support in [10], but it still has some limitations due to the use solely of shadows. First, not all of the shadows are cast by rooftops; there are other items (e.g., walls, fences, and high trees) that will cast shadows that are similar to the shadows cast by rooftops. This makes it difficult to exclude by simply using thickness of shadows or the distance from shadows to vegetation. Second, in some cases, there are no visible shadows in the aerial image.

It has been observed in [4], [5], and [11] that man-made objects rooftops possess quite regular shapes, which is a well-discriminating feature to distinguish rooftops from other objects. Inspired by the work of Cote and Saeedi [11], we propose to roughly classify the segments that are not shadows and vegetation into probable rooftops and probable nonrooftops depending on shape, shadows, and light direction. Since initial classification may assign the same label to disconnected regions, we relabel the GMM presegmentation result into 4-connected components based on their initial labels. We chose 4-connected components rather than 8-connected components to reduce the chance that two diagonally adjacent rooftops that touch at a corner would be accidentally merged into a single component. Then, the following steps are used to determine which blobs (components) belong to probable rooftops.

1) Size: Blobs with very small size or very large size are not likely to be rooftops; we empirically set the area range for rooftops to $[S_{\min}, S_{\max}]$.
2) Eccentricity: The eccentricity of each blob is defined as the ratio of the minor axis length and the major axis length of the ellipse that has the same second moments as the blob. The value ranges from zero to one, and we only select the blobs with eccentricity value greater than τ_e as probable rooftops to discard excessively elongated blobs.
3) Compactness: The compactness is defined as

$$c = 4A/P^2 \tag{2}$$

where A and P represent the area and the perimeter of blob, respectively; we only keep the blobs with compactness greater than τ_c as probable rooftops.

We want to emphasize that we do not require that each connected component is classified correctly in this step; the next steps of our algorithm are able to discard the mislabeled rooftops and recover the missing rooftops as long as the majority of the probable rooftops are correct. We will discuss the selection of these thresholds in Section IV-D.

We prune out the probable rooftops further, on condition that light direction is available. For each remaining blob that belongs to probable rooftops, we check its neighboring blobs in the light direction. If no shadows are found, we remove this blob from probable rooftops. Fig. 3 gives one example of removing nonprobable rooftops following the preceding steps.

4) Segmentation Based on HCRF: The final rooftop extraction can be formulated as a multilabel segmentation problem, and we adopt the commonly used CRF model to solve it effectively. The commonly used standard CRF model is expressed as the sum of unary and pairwise potentials as

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \tag{3}$$

where $\mathcal{V}$ denotes the set of all image pixels, $\mathcal{E}$ is the 8-neighboring pixel pairwise set connecting the pixels $i, j \in \mathcal{V}$, and x_i denotes the label taken by pixel i of the image; then, a segmentation over the image is defined as every possible assignment of $\mathbf{x}$.

The unary potentials $\psi_i(x_i)$ are the negative log of the likelihood of label x_i being assigned to pixel i, and the pairwise potential $\psi_{i,j}$ is a smoothness term to enable neighboring pixels to take the same label. In our case, we want to segment an aerial image into four classes: shadows, vegetation, rooftops, and unknown (see Fig. 4). Thus, the corresponding label set is $x_i \in \{0, 1, 2, 3\}$. Considering that we have already obtained an initial classification of the pixels through the above steps, we can estimate the unary potentials of each pixel from the classification result. We model the distribution of each class using GMM with different components (M); for each class, the parameters of GMM are initialized by fitting the pixels belonging to that class according to the initial classification

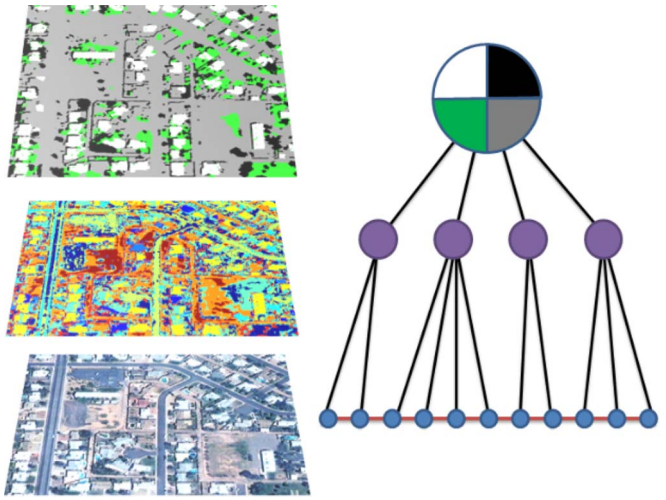

Fig. 4. Structure of a high-order CRF. Blue dots in the right image represent the pixels in the bottom left image, and purple dots represent segment formed by pixels shown in the middle image in the left. In our experiment, each pixel will be classified into four classes: shadows (black), vegetation (green), rooftops (black), and unknown (gray).

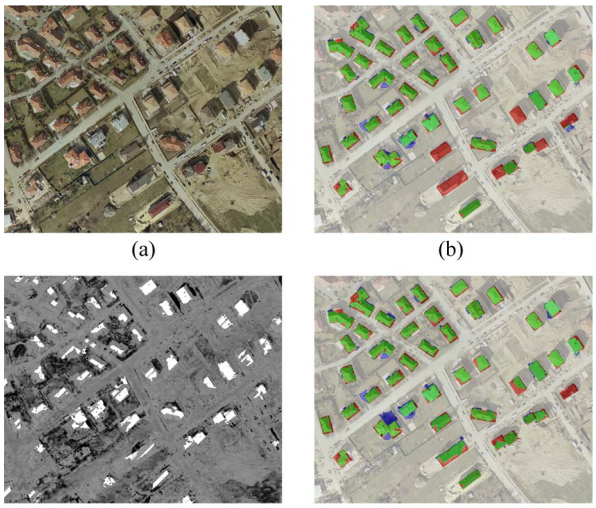

Fig. 5. Effectiveness of adding hard constraint from shadows. (a) Original image. (b) Result using ground truth data as the initial mask for multilabel graph cut. (c) Hard constraint obtained from our method. (d) Our result.

result. Then, unary potentials $\psi_i(x_i)$ are deduced from the GMM of each class.

However, this is not sufficient to capture all of the rooftops, particularly when there are rooftops that have similar color with the background region. An example is shown in Fig. 5(b); the rooftops marked by red color have a higher probability belonging to unknown class since they have the same appearance as the dominant component of unknown part, i.e., the roads. Thus, the following segmentation will miss those rooftops even if they have been correctly classified as probable rooftops in previous steps. Fig. 5(b) depicts such a failure case; we use ground truth to initialize the classification, but we still lose those rooftops in the final result. To overcome this problem, we propose to assign a high probability to those proposed rooftops that are adjacent to shadows on one or more directions. Specifically, for each probable rooftops $\mathcal{R}$ obtained from previous steps, if shadows

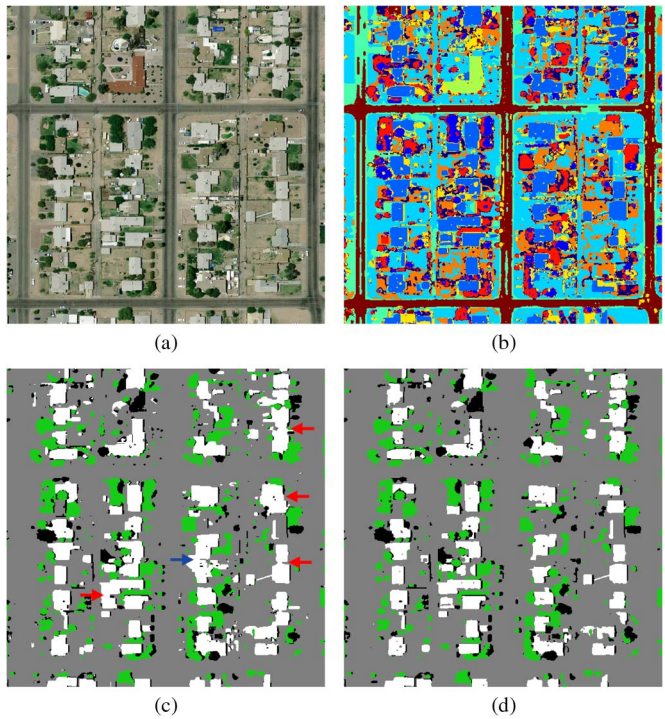

Fig. 6. Comparisons between common CRF and high-order CRF. (a) Original image. (b) Segments obtained from GMM presegmentation. (c) and (d) Final segmentation results of common CRF and HCRF, respectively. Rooftops, shadows, vegetation, and unknown are represented by white, black, green, and gray, respectively. Notice the differences indicated by arrows. Red arrows indicate the improvement by utilizing feature-based consistency potential, and blue arrow indicates the improvement by utilizing region-based consistency potential.

$\mathcal{S}$ can be found in the neighboring part of $\mathcal{R}$, we assume that pixels $\mathcal{R}_h$ should be part of rooftops, where

$$\mathcal{R}_h = \left(\mathcal{S} \oplus v_{(-L,d)}\right) \cap \mathcal{R} \tag{4}$$

where $\oplus$ denotes the morphological binary dilation operator, structure element $v_{(L,d_F)}$ is a line segment with one end at the origin of the structure element and the other end a distance d opposite the light direction L, and d is the maximum of the width and height of $\mathcal{R}$'s bounding box.

For the pairwise potentials, which encode dependencies between neighboring pixels, we typically define $\psi_{ij}(x_i, x_j)$ in the form of a contrast-sensitive Potts model, i.e.,

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0, & \text{if } x_i = x_j \\ \theta_\lambda \exp\left(-\theta_\beta \|I_i - I_j\|^2\right), & \text{otherwise} \end{cases} \tag{5}$$

where I_i and I_j denote the color vectors of pixels i and j, respectively; and x_i and x_j denote the labels taken by pixels i and j, respectively. For our application, we use Lab color space for illumination uniformity. θ_β is learned from the image using the way in [37]. θ_λ controls the balance between fitting term (unary potentials) and smoothness term (pairwise potentials).

The standard pairwise potentials in the CRF model tend to introduce oversmooth segmentation due to their inability to encode high-level structures. This limitation causes undesirable rooftops extraction result, particularly when rooftops are close to each other and the background between rooftops has low contrast. Notice the area indicated by red arrows in Fig. 6(c); multiple rooftops are merged into one single rooftop, which

reduces the object-level accuracy significantly. In order to alleviate this problem, we propose to utilize the new developed HCRF to avoid the ambiguities by incorporating high-level context. The HCRF model extends (3), by adding an additional term defined over higher order cliques, i.e.,

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}} \psi_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j) + \sum_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c). \quad (6)$$

Here, $\mathcal{C}$ represents a set of image segments obtained from Section III-A1, and ψ_c denotes high-order potentials defined on segment c. Fig. 4 illustrates how to construct the HCRF model from pixels and segments. Notice that the aim of our multilabel segmentation is to extract rooftops as accurately as possible; we will explain how to specialize the definition of high-order potentials for this goal.

- *Region-Based Consistency Potential*. This potential enforces the pixels in one segment c to take the same label. Hence, high-level structure information, captured by the presegmentation, can be taken into account during the optimization. The region indicated by blue arrows in Fig. 6(c) illustrates how we can benefit from high-level structures. As shown in Fig. 6(c), the mislabeled rooftops (point by the blue arrow) are a small portion of one single segment [see Fig. 6(b)]. However, the remaining parts of this segment are labeled as nonrooftops, leading to inconstant labels in one segment. The region-based consistency potential is designed to penalize for inconsistency of labels in one segment, so that mislabeled rooftops can be effectively removed, as shown in Fig. 6(d). We define the region-based consistency potential $\psi_c^r(\mathbf{x}_c)$ as

$$\psi_c^r(\mathbf{x}_c) = \begin{cases} \lambda_k^r, & \text{if } x_i = l_k, \quad \forall i \in c \\ \lambda_{\max}^r, & \text{otherwise} \end{cases} \quad (7)$$

  where $\lambda_k^r \le \lambda_{\max}^r$, and $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$ form the label set taken by all of the pixels. In our case, $\mathcal{L} = \{0, 1, 2, 3\}$.
- *Feature-Based Consistency Potential*. We assume that the quality of each blob as a rooftop can be evaluated based on several features. It has been discussed in Section III that shape of blobs is a discriminating feature for identifying rooftops. Suppression of very small segment is particularly useful in removing those tiny and thin pieces between neighboring rooftops. These are normally detected as rooftops due to the shrink bias, caused by the smoothing term in the CRF model, as shown in the region indicated by red arrows in Fig. 6(c). We also evaluate the quality using the intensity variance in each blob, since rooftops tend to not have too much intensity variation. Combining these features, the feature-based consistency potential $\psi_c^f(\mathbf{x}_c)$ is defined as

$$\psi_c^f(\mathbf{x}_c) = \begin{cases} 0, & \text{if } x_i = \{0,1,3\}, \quad \forall i \in c \\ \lambda_{\max}^f, & \text{if } x_i = 2, \quad \forall i \in c \text{ and } |c| \le \theta_s \\ \theta_h \exp\left(-\theta_\alpha \sigma_c^2\right) & \\ \quad \cdot \lambda_{\max}^f, & \text{if } x_i = 2, \quad \forall i \in c \text{ and } |c| > \theta_s \\ \lambda_{\max}^f, & \text{otherwise} \end{cases} \quad (8)$$

  where $|c|$ is the area of segment c, θ_s denotes the area threshold that segments with area less than it will be given a high cost $\lambda_{\max}^f$ to be rooftops, σ_c is the standard deviation of the intensity of pixels inside c, and its impact is controlled using weight coefficients θ_h and θ_α.

While it is hard to ensure that each of the segments matches the objects exactly, we adopt the robust form of the high-order potentials [38] to allow some pixels in one segment to take different labels. Then, the final high-order potentials are written as

$$\psi_c(\mathbf{x}_c) = \min\left\{\min_{k\in\mathcal{L}}\left((|c| - n_k(\mathbf{x}_c)) \cdot \frac{\lambda_{\max} - \lambda_k}{Q} + \lambda_k\right), \lambda_{\max}\right\} \quad (9)$$

where $n_k(\mathbf{x}_c)$ denotes the number of pixels in segment c, which takes label k, Q is a truncation parameter, which controls how many pixels in one segment can take different labels and satisfies the constraint $2Q < |c|$, $\lambda_{\max} = \max(\lambda_{\max}^r, \lambda_{\max}^f)$, and $\lambda_k = \max(\psi_c^r(\mathbf{x}_c), \psi_c^f(\mathbf{x}_c))$. During our experiments, we set $Q = 0.1|c|$ and $\lambda_{\max}^r = \lambda_{\max}^f = 2 \cdot \max\{\psi_i(x_i)\}$, where $\psi_i(x_i)$ is defined in (6). We finally solve (9) using the implementation of the expansion and swap move algorithms described in [38] for its excellent computational performance.

IV. Evaluation

We implemented our algorithm in Python, first testing our approach on the SZTAKI-INRIA Building Detection Benchmark [12] and our own data set. Then, we analyze the sensitivity of important parameters used in our algorithm and discuss the limitations of the proposed method.

A. Evaluation on the SZTAKI-INRIA Benchmark

The SZTAKI-INRIA Benchmark [12] is an excellent resource for benchmarking building extraction algorithms. It contains the rectangular footprints of 665 buildings in nine aerial or satellite images taken from Budapest and Szada (both in Hungary), Manchester (U.K.), Bodensee (Germany), and Normandy and Cot d'Azur (both in France), as well as manually annotated ground truth data. Two data sets (Budapest and Szada) are aerial images, and the remaining four data sets are satellite images acquired from Google Earth. All of the images contain only RGB information.

We perform quantitative evaluation both at object and pixel levels. We use the common measures of precision, i.e., P, recall, i.e., R, and the F-score, i.e., F_1, to measure pixel level accuracy, where

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (11)$$

$$F_1 = \frac{2PR}{P + R}. \quad (12)$$

Here, TP stands for true positives and refers to the number of pixels assigned as rooftop in both ground truth and segmentation result. FP stands for false positives and refers to the

TABLE I
NUMERICAL OBJECT-LEVEL AND PIXEL-LEVEL COMPARISONS BETWEEN STATE-OF-THE-ART BUILDING DETECTION METHODS AND THE PROPOSED METHOD (HCRF) WITH THE BEST RESULTS IN BOLD. FOR THE EVALUATION OF THE SIFT METHOD, THE MANCHESTER DATA SET IS IGNORED DUE TO WEAK PERFORMANCE

Data Set		Object level performance												Pixel level performance									
		SIFT		Gabor		EV		SM		MPP		Prop. HCRF		EV			SM			MPP			Prop. HCRF		
Name	#obj	MO	FO	MO	FO	MO	FO	MO	FO	MO	FO	MO	FO	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BUDAPEST	41	20	10	8	17	11	5	9	1	2	4	1	0	0.73	0.46	0.56	0.84	0.61	0.70	0.82	0.71	0.76	0.90	0.75	**0.81**
SZADA	57	17	26	17	23	10	18	11	5	4	1	2	3	0.61	0.62	0.61	0.79	0.71	0.74	0.93	0.75	0.83	0.85	0.86	**0.85**
COTE D'AZUR	123	55	9	12	24	14	20	20	25	5	4	3	4	0.73	0.51	0.60	0.75	0.61	0.67	0.83	0.69	0.75	0.75	0.81	**0.77**
BODENSEE	80	34	9	32	8	11	13	18	15	7	6	4	3	0.56	0.30	0.39	0.59	0.41	0.48	0.73	0.51	0.60	0.84	0.76	**0.79**
NORMANDY	152	69	14	24	14	18	32	30	58	18	1	16	7	0.60	0.32	0.41	0.62	0.55	0.58	0.78	0.60	0.67	0.79	0.67	**0.72**
MANCHESTER	171	NA	NA	53	85	46	17	53	42	19	6	20	3	0.64	0.38	0.47	0.60	0.56	0.57	0.86	0.63	0.72	0.82	0.67	**0.73**
Overall F-score		0.662		0.751		0.827		0.771		0.936		**0.948**		0.517			0.631			0.726			**0.786**		

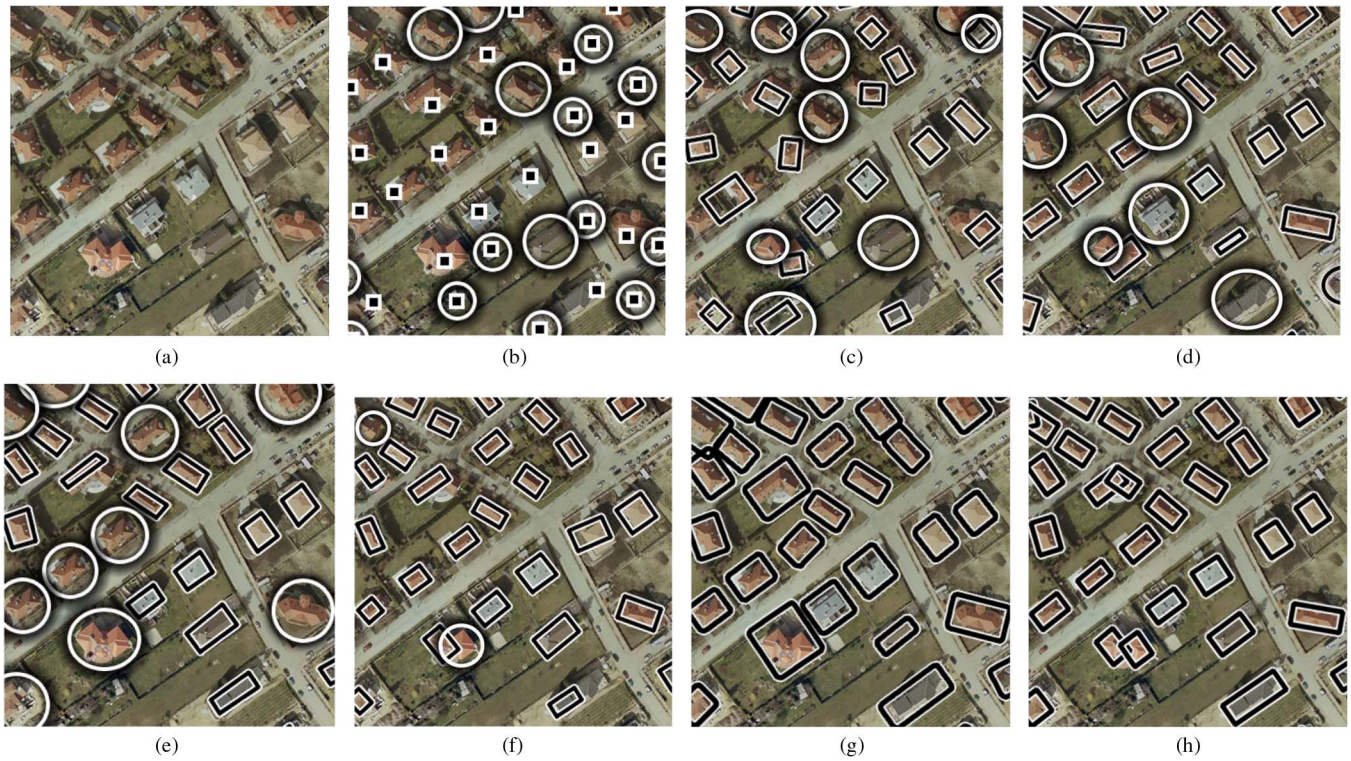

Fig. 7. Comparisons between the proposed method and the state-of-the-art methods. The results of state-of-the-art methods are from [12]. White circles represent missing or false objects, and black rectangles represent true positive objects. For the Gabor method, only the building center is extracted. (a) Original image. (b) Gabor [16]. (c) EV [25]. (d) SM [13]. (e) MRF [17]. (f) MPP [12]. (g) Proposed HCRF. (h) Ground truth.

number of pixels assigned as rooftop in result but not in ground truth. FN stands for false negatives and refers to the number of pixels assigned as rooftop in ground truth but not in result. The F-score F_1 captures both precision and recall into a single metric that gives each an equal importance. We evaluate the object-level performance by counting the missing and falsely labeled rooftops (MO and FO, respectively). Then, we use the same formula to calculate the F-score at object level [10].

Table I lists the numerical object-level and pixel-level comparisons with SIFT [15], Gabor [16], MRF [17], EV [25], SM [13], MPP [12], and the proposed HCRF method. We show that our method outperforms the best of the state-of-the-art methods (MPP) by 6% at pixel level and 1% percent at object level.

Fig. 7 gives qualitative comparison results on the BUDAPEST data set. The Gabor [16] method fails to capture dark buildings due to the false gradient directions. Without training data, the EV [25] method can only extract red rooftops and lacks the ability to deal with the color diversity of rooftops in different regions, resulting in high MO. The SM [13] technique obtains relatively high precision, at the cost of a rather low recall, as shown in Fig. 7(d). This is because, if one building is missing during the filtering of initial segmentation, there is no way to then recover it in the following steps.

We then discuss random-field-based methods, i.e., MRF [17] and MPP [12]. Compared with the deterministic decision rules used in the preceding methods, the probabilistic model is able to describe dependencies between multiple hypotheses and find a globally optimal segmentation based on specific data likelihood. However, since the MRF technique constructs hypothesis graph based on the edges of buildings, which is sensitive to the resolution and noise of the image, from Fig. 7(e), we can see that they still miss 10 out of 41 buildings. We want to point out that local descriptors such as edges and corners are more sensitive to variation on resolution, image quality, and noise due to the absence of high-level information. By combining features at different levels, this method improves both MO and FO a lot,

as shown in Fig. 7(f). However, they also run the stochastic object extraction process only at object level, which impairs accuracy at pixel level, as listed in Table I. Another limitation of MPP is the requirement of ground truth data for each image to determine the roof color filtering threshold.

Our method surpasses all of the above methods, as shown in Fig. 7(g). We accomplish this through several key strategies. First, we use GMM to perform the presegmentation and extract shadows and vegetation based on the presegmentation. Despite its simplicity, experiments reveal that it is robust to give reasonable presegmentation result on different data sets. Detailed analysis can be found in Section IV-D. Second, we employ HCRF to achieve a global optimal segmentation. The results confirm that the combination of pixel-level feature and segment-level feature might solve the dilemma between pixel-level accuracy and object-level accuracy efficiently.

One more challenging example is shown in Fig. 8(l). The test region contains buildings with complex shapes and varying sizes; the proposed method still gives high-quality extraction result with 0.92 precision, 0.82 recall, and F-score of 0.87.

B. Comparison With the Standard CRF Method

We compare our method with the automated detection method proposed in [1], which uses standard CRF model at pixel level. Since their method requires NIR band, we use another data set taken from Tempe, AZ, USA, as shown in Fig. 9. The 512×512 image has RGB, NIR information, and a resolution of 0.6 m per pixel. The method of Ok [1] fails to capture most of the rooftops with 0.65 precision, 0.41 recall, and F-score of 0.51 (upper row) and 0.85 precision, 0.39 recall, and F-score of 0.54 (bottom row), as shown in Fig. 9(d). Our approach achieves a much better result, as shown in Fig. 9(e) with 0.81 precision, 0.65 recall, and F-score of 0.72 (upper row) and 0.89 precision, 0.68 recall, and F-score of 0.77 (bottom row). The method of Ok deduces directional constraint for rooftops from shadows in their first step. They only will use the most reliable shadows in order to reduce the false positive rooftops. This, however, will miss a lot of rooftops, leading to a low recall. Some of the missing rooftops are hard to recover, even through a global multilabel graph cut in the second step. Our method obtains the initial candidate rooftops by classifying the presegmentation results. Shadows only provide supplementary evidence to filtering the initial rooftops furthermore. This noticeably improves the recall, as shown in Fig. 9(c), the downside being more false positive rooftops. We show that the proposed HCRF, which incorporates the features both from pixel and segment levels, further eliminates most mislabeled rooftops. This is achieved while maintaining high precision and recall, as shown in Fig. 9(e).

To better illustrate the advantage of the HCRF method, we give the result of our method only using standard CRF in Fig. 9(b). We list the performance values of applying our methods on the benchmark only using standard CRF in Table II. Comparing with HCRF, standard CRF generates oversmoothed results with higher recall but at the cost of dramatic drop in precision.

C. Comparison With the Nonshadow-Based Method

The proposed method has the advantage that it can produce competitive result even without using shadows. We demonstrate the robustness of our method and compare with the state-of-the-art nonshadow-based method [11] in Fig. 10. The data are an aerial image taken from San Diego, CA, USA, with a resolution of 0.23 m per pixel. Affected by clouds, it is hard to detect distinct shadows in the image. We use source code (provided by the author) to produce the result of Cote and Saeedi, using the parameters suggested in their papers. For fairness, we remove rooftops touch image border from our final result as they did. Since the boundaries of rooftops are blurred in this image, we do not consider the pixels lying within 2 pixels around the boundary of ground truth when calculating the numerical performance. Our method surpasses the method of Cote and Saeedi [11], with 0.94 precision, 0.86 recall, and F-score of 0.90 (upper row) and 0.90 precision, 0.90 recall, and F-score of 0.90 (bottom row) compared with 0.71 precision, 0.61 recall, and F-score of 0.65 and 0.81 precision, 0.79 recall, and F-score of 0.80, respectively. This is not surprising, since the final CRF-based segmentation in our method can help to correct the misclassified presegmentation. The method of Cote and Saeedi, on the other hand, suffers from presegmentation having similar shape with rooftops, but which do not belong to any rooftops.

To reveal how shadows information affects the overall performance, we run our method without using shadows information and give the numerical result in Table II. As mentioned in Fig. 5, using shadows information is important to reduce the number of missing rooftops. According to the performance values in Table II, we observed that using shadows information was able to improve the overall performances by 2%.

D. Parameter Settings

Table III lists the default settings of the parameters used in the proposed method. Fig. 11(a)–(d) illustrates the effects of choosing different values for extracting probable rooftops. We select the default values of these thresholds through analyzing statistics on real-world data. As we point out in Section III-A3, the purpose of introducing these thresholds is to rule out the segments that most unlikely belong to rooftops. As long as most of the rooftops are kept during this step, the proposed method achieves high performance. As proven in these figures, when the parameters vary around the default values, there is only slight change in the final F-score.

We then discuss the sensitivity of the component number M of the GMM used in Section III-A1. The effect of parameter variation on the data set MANCHESTER image is shown in Fig. 11(e). As we increase M from 5 to 15, the precision keeps increasing since the boundaries between different objects become more accurate. However, if M becomes too large, it leads to oversegmentation of the image, breaking one rooftop into a multitude of small pieces. The result being a decreased recall. A large M also increases the cost of time spent on segmentation and final HCRF optimization. During our experiment, we set $M = 10$ for all of the data sets.

Fig. 8. Gallery of our results on the benchmark data set and our own data set. The images are taken from the Bodensee, Normandy, Cot d'Azur, and Manchester data sets, respectively. The last image is taken from Casa Grande area. Correct results (TP) are shown in green, false positives are shown in blue, and false negatives are shown in red.

Another important parameter is the weighting coefficients θ_λ, as defined in (5), which control the smoothness of segmentation. Fig. 11(f) shows the segmentation accuracy of both standard CRF and high-order CRF at different smoothing weights θ_λ. The standard CRF is much more sensitive to the changes of smoothing weight; when θ_λ increases, the segmentation tends to be oversmooth. This improves the recall, but at the cost of significant drop in precision. In contrast,

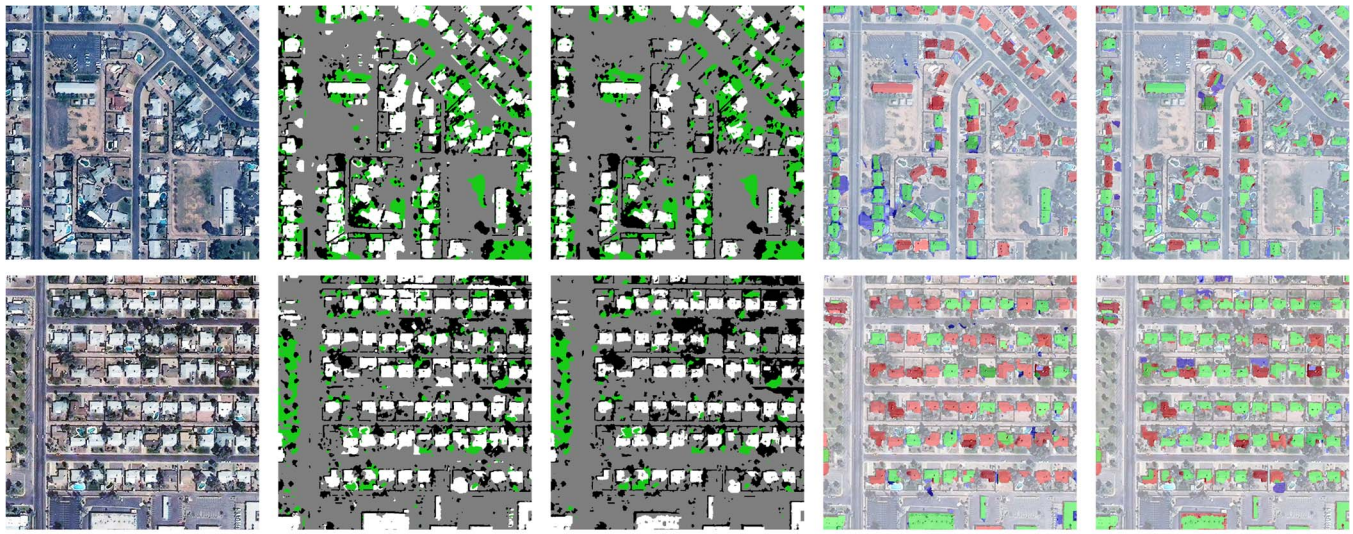

Fig. 9. Comparisons with the standard CRF method [1]. (a) Original image. (b) and (c) Results of our method using standard CRF model and HCRF model, where rooftops, shadows, vegetation, and unknown are represented by white, black, green, and gray colors, respectively. (d) Buildings extracted using the method in [1]. (e) Buildings extracted using our HCRF method. Correct results (TP) are shown in green, false positives are shown in blue, and false negatives are shown in red.

TABLE II
NUMERICAL OBJECT LEVEL AND PIXEL LEVEL OF APPLYING OUR APPROACH USING STANDARD CRF (PROP.CRF), APPLYING OUR APPROACH WITHOUT SHADOWS INFORMATION (PROP.HCRF W/O SHADOWS), AND THE PROPOSED METHOD (HCRF) WITH THE BEST RESULTS IN BOLD

Data set		Object level performance						Pixel level performance								
		Prop.CRF		Prop.HCRF w/o Shadows		Prop.HCRF		Prop.CRF			Prop.HCRF w/o Shadows			Prop.HCRF		
Name	#obj	MO	FO	MO	FO	MO	FO	P	R	F_1	P	R	F_1	P	R	F_1
BUDAPEST	41	1	0	5	0	1	0	0.84	0.78	0.80	0.91	0.69	0.78	0.90	0.75	**0.81**
SZADA	57	2	3	3	3	2	3	0.80	0.87	0.83	0.85	0.80	0.83	0.85	0.86	**0.85**
COTE D'AZUR	123	3	5	4	4	3	4	0.68	0.84	0.75	0.74	0.79	0.76	0.75	0.81	**0.77**
BODENSEE	80	4	3	6	2	4	3	0.81	0.77	0.78	0.85	0.70	0.76	0.84	0.76	**0.79**
NORMANDY	152	14	11	19	7	16	7	0.64	0.77	0.69	0.79	0.61	0.68	0.79	0.67	**0.72**
MANCHESTER	171	18	8	27	2	20	3	0.63	0.79	0.70	0.85	0.61	0.71	0.82	0.67	**0.73**
Overall F-score		0.941		0.931		**0.948**		0.767			0.765			**0.786**		

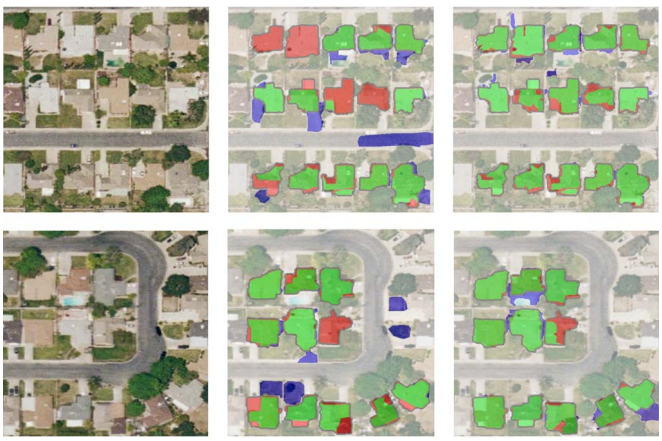

Fig. 10. Comparisons with the nonshadow-based method [11]. (a) Original image. (b) Buildings extracted using the method in [11]. (c) Our result. Correct results (TP) are shown in green, false positives are shown in blue, and false negatives are shown in red.

TABLE III
PARAMETER SETTINGS FOR THE PROPOSED APPROACH

Parameter	Value
Minimum Rooftop Size (S_{min})	$10m^2$
Maximum Rooftop Size (S_{max})	$1000m^2$
Eccentricity Threshold (τ_e)	0.175
Compactness Threshold (τ_c)	0.15
GMM Component Number in Section III (M)	10
Weighting Coefficients in Section III-A4 (θ_λ)	2
Higher Order Potential Coefficient (θ_h)	0.5
Higher Order Potential Coefficient (θ_α)	12
Higher Order Potential Coefficient (θ_s)	$10m^2$

with higher term in the CRF model, our approach exhibits high robustness, and the performance of our high-order CRF remains stable when θ_λ varies from 1 to 10. This is because the higher term constraints efficiently alleviate the oversmooth segmentation among rooftops close to each other, as explained in Section III-A4. In our experiment, we use a constant $\theta_\lambda = 2$ in (5).

For the coefficients used for calculating the higher order potential, as defined in (7)–(9), we use $\theta_s = 10$ m^2, $\theta_h = 0.5$, and $\theta_\alpha = 12$. $\theta_s = 10$ m^2 is used to suppress very small segments between neighboring rooftops caused by the smoothing term; as shown in Fig. 11(i), noticeable improvement is observed when θ_s varies from 0 to 2 m^2. When θ_s keeps growing, the

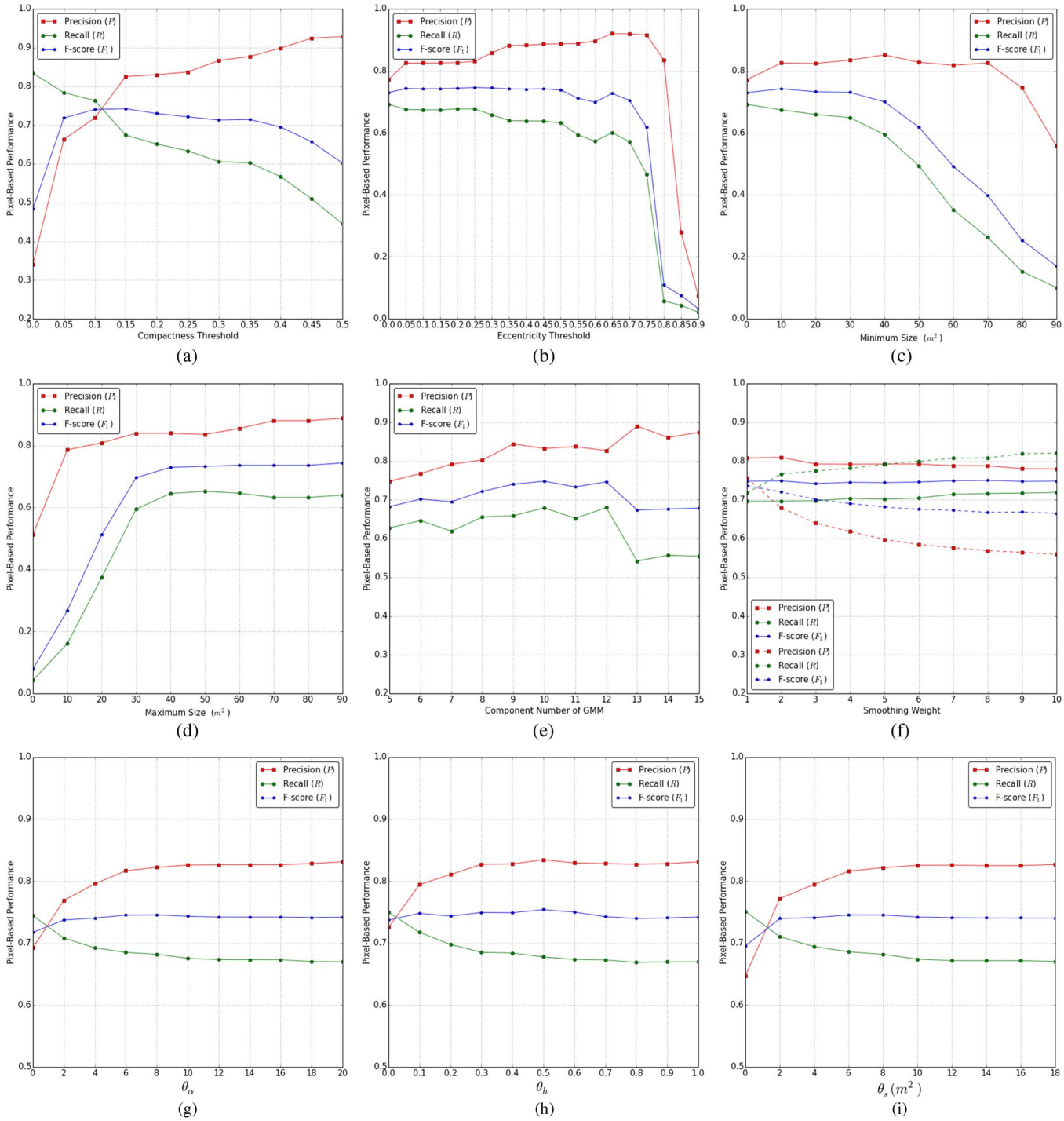

Fig. 11. Pixel-based performance in the case of different parameter settings. (a) Effects of compactness threshold (τ_c). (b) Effects of eccentricity threshold (τ_e). (c) and (d) Effects of minimum and maximum rooftop sizes $(S_{\min}, S_{\max})$. (e) Effects of GMM component number (M). (f) Effects of weighting coefficients (θ_λ); solid line denotes HCRF, and dash line denotes CRF. (g)–(i) Effects of coefficients in high-order potential $(\theta_\alpha, \theta_h, \theta_s)$.

performance values stay stable and consistent. This is because the eliminated regions are relatively small; most of the areas are around 0–2 m^2. In our experiments, we choose the optimal value of θ_s the same as $S_{\min}$ for both robustness and performance. Considering the parameters θ_α and θ_h, which are used to control the effect of higher order potential, we follow the guidelines in [38], using $\theta_\alpha = 12$ and $\theta_h = 0.5$. Although these values are learned from close-range data sets, we found in our experiments that they are also valid for our data sets and give satisfactory results. We further explore using other values; however, simply changing one parameter affects the overall performance less than 1%, as shown in Fig. 11(g) and (h). We plan to perform grid search for combination of these parameters to find the optimal parameter setting in the future.

Finally, regarding the GMM component numbers used for calculating the unary potential in Section III-A4, as revealed in

[1], we set the optimal component numbers of shadows, vegetation, rooftops, and unknown to 2, 2, 8, and 8, respectively.

E. Limitations

There are several limitations to our approach for rooftop extraction. First, although the proposed shadows and vegetation extraction method works well on most of the data sets, it fails to capture all of the shadows and vegetation correctly in several cases. If the illumination within shadow region is not constant due to changes in reflectance, some of the shadows will be mislabeled as rooftops. Such a failure case is shown in Fig. 8(f). Considering that some of the shadows have the same intensity as the dark rooftops in this case, additional information besides color should be taken into account to detect the shadows correctly. We plan to explore a new way to solve this problem in the future. Very dark objects such as road and water regions would be mislabeled as shadows using the method in Section III-A2, which will also introduce errors when adding hard constraint from shadows. In complicated urban regions, rooftops vary in height, and the shadows of one rooftop may be cast onto another roof, which will cause incorrect holes in the extracted rooftops. We extract the vegetation based on the greenness; thus, if there are green rooftops in the image, our method will mislabel them as vegetation in Section III-A2 and cannot recover them in the final extracted rooftops.

Second, the presegmentation fails to represent the shape of rooftops in the case of large noise, low contrast, and similar color with background region. As shown in Fig. 8(h), several rooftops are missing due to the low image quality. Those rooftops are identified as nonrooftops in Section III-A3 because the GMM segmentation generates wrong segments for them, and the final HCRF-based segmentation fails to recover them either since they have the similar color with the roads. We notice that most of the methods listed in Table I give weak performance on this data set.

V. Conclusion and Future Work

In this paper, we have proposed a new framework for building extraction in remote sensing images. Our method incorporates pixel- and segment-level information for the identification of rooftops. Based on the segments obtained from an unsupervised segmentation, the proposed method automatically extracts vegetation, shadows, and probable rooftops. Then, an HCRF segmentation is used to achieve accurate rooftop extraction, by exploiting color features at the pixel level and region consistency and shape features at the segment level. We test our method on a variety of data sets, and the results reveal that the proposed HCRF segmentation improves the performance of rooftop extraction, both at pixel and object levels. Furthermore, the framework is efficient to deal with rooftops of complex shapes, without requirement of user prelabeled ground truth data. In the future, we plan to explore the use of associative hierarchical CRFs to further improve the accuracy or to incorporate spatial information during the initial GMM segmentation to better deal with noise and blurry in the image. We also plan to use our segmentation results to guide the remote sensing image compression.

Acknowledgment

The authors would like to thank M. Cote and P. Saeedi for providing the source code for [11] to compare against their method. They would also like to thank A. O. Ok for providing us with output to compare against the proposed method. In addition, they would also like to thank the Decision Center for a Desert City for providing the imagery used in Fig. 9.

References

[1] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogramm. Remote Sens.*, vol. 86, pp. 21–40, Dec. 2013.
[2] C. Brenner, "Building reconstruction from images and laser scanning," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 6, no. 3/4, pp. 187–198, Mar. 2005.
[3] C. Ünsalan and K. L. Boyer, "A system to detect houses and residential street networks in multispectral satellite images," *Comput. Vis. Image Understand.*, vol. 98, no. 3, pp. 423–461, Jun. 2005.
[4] Z. Liu, J. Wang, and W. P. Liu, "Building extraction from high resolution imagery based on multi-scale object oriented classification and probabilistic Hough transform," in *Proc. IEEE IGARSS*, Jul. 2005, vol. 4, pp. 2250–2253.
[5] S. Cui, Q. Yan, and P. Reinartz, "Graph search and its application in building extraction from high resolution remote sensing imagery," in *Search Algorithms and Applications*. Shanghai, China: InTech, 2011.
[6] Z. Liu, S. Cui, and Q. Yan, "Building extraction from high resolution satellite imagery based on multi-scale image segmentation and model matching," in *Proc. Int. Workshop EORSA*, Jun. 2008, pp. 1–7.
[7] D. Marr, "Early processing of visual information," *Philos. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 275, no. 942, pp. 483–519, Oct. 1976.
[8] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 10–17.
[9] H. Akcay and S. Aksoy, "Building detection using directional spatial constraints," in *Proc. IEEE IGARSS*, 2010, pp. 1932–1935.
[10] A. Ok, C. Senaras, and B. Yuksel, "Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1701–1717, Mar. 2013.
[11] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2013.
[12] C. Benedek, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth–death dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 33–50, Jan. 2012.
[13] S. Müller and D. W. Zaum, "Robust building detection in aerial images," in *Proc. Int. Arch. Photogramm. Remote Sens.*, 2005, pp. 143–148.
[14] L. Martinez Fonte, S. Gautama, W. Philips, and W. Goeman, "Evaluating corner detectors for the extraction of man-made structures in urban areas," in *Proc. IEEE IGARSS*, 2005, vol. 1–8, pp. 237–240.
[15] B. Sirmacek and C. Ünsalan, "Urban-area and building detection using sift keypoints and graph theory," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1156–1167, Apr. 2009.
[16] B. Sirmacek and C. Ünsalan, "A probabilistic framework to detect buildings in aerial and satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 211–221, Jan. 2011.
[17] A. Katartzis and H. Sahli, "A stochastic framework for the identification of building rooftops using a single remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 259–271, Jan. 2008.
[18] M. S. Nosrati and P. Saeedi, "A novel approach for polygonal rooftop detection in satellite/aerial imageries," in *Proc. 16th IEEE Int. Conf. Image Process.*, 2009, pp. 1689–1692.
[19] R. B. Irvin and D. M. McKeownJr., "Methods for exploiting the relationship between buildings and their shadows in aerial imagery," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 6, pp. 1564–1575, Nov./Dec. 1989.
[20] C. Lin, A. Huertas, and R. Nevatia, "Detection of buildings using perceptual grouping and shadows," in *Proc. IEEE CVPR*, 1994, pp. 62–69.
[21] C. Lin and R. Nevatia, "Building detection and description from a single intensity image," *Comput. Vis. Image Understand.*, vol. 72, no. 2, pp. 101–121, Nov. 1998.
[22] T. Kim, T. Javzandulam, and T.-Y. Lee, "Semiautomatic reconstruction of building height and footprints from single satellite images," in *Proc. IEEE IGARSS*, 2007, pp. 4737–4740.

[23] T. Kim and J.-P. Muller, "Development of a graph-based approach for building detection," *Image Vis. Comput.*, vol. 17, no. 1, pp. 3–14, Jan. 1999.
[24] Y.-T. Liow and T. Pavlidis, "Use of shadows for extracting buildings in aerial images," *Comput. Vis. Graph. Image Process.*, vol. 49, no. 2, pp. 242–277, Feb. 1990.
[25] B. Sirmacek and C. Ünsalan, "Building detection from aerial images using invariant color features and shadow information," in *Proc. 23rd ISCIS*, 2008, pp. 1–5.
[26] J. Femiani and E. Li, "Graph cuts to combine multiple sources for feature extraction," in *Proc. IMAGE*, Dayton, OH, USA, 2014.
[27] J. Femiani, E. Li, A. Razdan, and P. Wonka, "Shadow-based rooftop segmentation in visible band images," *IEEE Sel. Topics Appl. Earth Observ. Remote Sens.*, to be published.
[28] J. Wegner, U. Soergel, and B. Rosenhahn, "Segment-based building detection with conditional random fields," in *Proc. JURSE*, Apr. 2011, pp. 205–208.
[29] D. Zoran and Y. Weiss, "Natural images, Gaussian mixtures and dead leaves," in *Proc. NIPS*, 2012, pp. 1745–1753.
[30] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," in *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.
[31] M. Tkalcic and J. F. Tasic, "Colour spaces: Perceptual, historical and applicational background," in *Proc. IEEE EUROCON Comput. Tool Region 8*, 2003, vol. 1, pp. 304–308.
[32] J. A. Bilmes *et al.*, *A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, vol. 4. Berkeley, CA, USA: International Computer Science Institute, 1998, p. 126.
[33] K.-L. Chung, Y.-R. Lin, and Y.-H. Huang, "Efficient shadow detection of color aerial images based on successive thresholding scheme," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 2, pp. 671–682, Feb. 2009.
[34] V. J. Tsai, "A comparative study on shadow compensation of color aerial images in invariant color models," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1661–1671, Jun. 2006.
[35] N. Shorter and T. Kasparis, "Automatic vegetation identification and building detection from a single nadir aerial image," *Remote Sens.*, vol. 1, no. 4, pp. 731–757, Oct. 2009.
[36] M. Chikr El-Mezouar, N. Taleb, K. Kpalma, and J. Ronsin, "Vegetation extraction from IKONOS imagery using high spatial resolution index," *J. Appl. Remote Sens.*, vol. 5, no. 1, 2011, Art. ID. 053543.
[37] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *Proc. 8th IEEE ICCV*, 2001, vol. 1, pp. 105–112.
[38] P. Kohli, L. Ladický, and P. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 302–324, May 2009.

Er Li (M'13) received the B.S. degree in automation from Wuhan University, Wuhan, China, in 2007 and the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

From 2012 to 2013, he was a Postdoctoral Researcher with the Institute of Software, Chinese Academy of Sciences, Beijing. He is currently a Postdoctoral Researcher with the Department of Engineering and Computing Systems, Arizona State University, Mesa, AZ, USA. His research interests include image analysis, computer vision, and computer graphics.

John Femiani (M'07) received the Ph.D. degree in computer science from Arizona State University, Phoenix, AZ, USA, in 2009.

He is currently an Assistant Professor with the Department of Engineering, Arizona State University, Mesa, AZ. His research interests include topics in computer graphics, visualization, computer vision, remote sensing, and image processing.

Shibiao Xu (M'15) received the B.S. degree in information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2009 and the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2014.

He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests include vision understanding, 3-D reconstruction, and image-based modeling.

Xiaopeng Zhang (M'11) received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1984 and 1987, respectively, and the Ph.D. degree in computer science from the Institute of Software, Chinese Academy of Sciences, Beijing, China, in 1999.

He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His main research interests are computer graphics and computer vision.

Peter Wonka (M'05) received the Ph.D. degree in computer science and the M.S. degree in urban planning from the Technical University of Vienna, Vienna, Austria, in 2001 and 2002, respectively.

He was a Postdoctoral Researcher with the Georgia Institute of Technology, Atlanta, GA, USA, for two years. He is currently a Professor with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, and also an Associate Professor with Arizona State University, Tempe, AZ, USA. His research interests include topics in computer graphics, visualization, computer vision, remote sensing, image processing, and machine learning.