

Automatic Detection of Compound Structures by Joint Selection of Region Groups From a Hierarchical Segmentation

H. Gökhan Akçay and Selim Aksoy, *Senior Member, IEEE*

Abstract—A challenging problem in remote sensing image analysis is the detection of heterogeneous compound structures such as different types of residential, industrial, and agricultural areas that are composed of spatial arrangements of simple primitive objects such as buildings and trees. We describe a generic method for the modeling and detection of compound structures that involve arrangements of an unknown number of primitives in large scenes. The modeling process starts with a single example structure, considers the primitive objects as random variables, builds a contextual model of their arrangements using a Markov random field, and learns the parameters of this model via sampling from the corresponding maximum entropy distribution. The detection task is formulated as the selection of multiple subsets of candidate regions from a hierarchical segmentation where each set of selected regions constitutes an instance of the example compound structure. The combinatorial selection problem is solved by the joint sampling of groups of regions by maximizing the likelihood of their individual appearances and relative spatial arrangements. Experiments using very high spatial resolution images show that the proposed method can effectively localize an unknown number of instances of different compound structures that cannot be detected by using spectral and shape features alone.

Index Terms—Context modeling, Gibbs sampling, Markov random field (MRF), maximum entropy distribution, object detection, spatial relationships, Swendsen–Wang sampling.

I. INTRODUCTION

THE increasing spatial and spectral resolutions of the images acquired from new-generation satellites have improved the capability to capture additional details about the objects of interest and have increased the feasibility of new applications that rely on the effective identification of these objects. A common approach to object-based image classification and object recognition is to assume the existence of homogeneous regions that can be modeled with spectral or shape features alone. However, as the spatial resolution increases, such homogeneous regions often correspond to very small details. Consequently, a new requirement for semantic image understanding has become the modeling and identification of image regions that are intrinsically heterogeneous. Examples of



Fig. 1. Examples of compound structures in WorldView-2 images. Each 150×150 pixel window includes one or more examples for residential, industrial, and agricultural structures composed of various spatial arrangements of primitives (buildings and trees) with different color and shape characteristics.

such regions, also called *compound structures*, include different types of residential, industrial, and agricultural areas that are composed of spatial arrangements of simple primitive objects such as buildings and trees [1]–[3] as shown in Fig. 1. However, the detection of these structures is a challenging problem because there is no single color, shape, or texture feature that can effectively model their appearances.

One of the most common alternatives is to use a window-based approach where the image is divided into tiles and these tiles are classified according to their features. The bag of words (BoW) model has been popular in recent years for modeling the tile content. First, visual words are formed by quantizing local features. Then, each tile is described by the frequency of these words and is classified [4]–[6] or retrieved [7], [8]. The main problem in the BoW representation is that it does not consider spatial arrangements that can be very crucial for many types of compound structures. In other words, BoW is a first-order model that primitives contribute independently of their position and independently from each other.

In an attempt to exploit spatial information, Vaduva *et al.* [9] modeled relative positions between objects by extracting object pair signatures as words that characterize the tiles. However, the tile-based modeling still enforces artificial boundaries on the image. Segmentation algorithms can produce flexible boundaries and promise to be adaptive to the image content. For example, Kurtz *et al.* [10] extracted heterogeneous objects in multiple levels of details where the segmentation in the high-resolution image was provided by clustering the segmentation in a lower resolution image. Gaetano *et al.* [11] performed

Manuscript received July 26, 2015; revised October 26, 2015; accepted December 9, 2015. Date of publication February 8, 2016; date of current version April 27, 2016. This work was supported in part by the TUBITAK Grant 109E193.

The authors are with the Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey (e-mail: akcay@cs.bilkent.edu.tr; saksoy@cs.bilkent.edu.tr).

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.2016.2519245

hierarchical texture segmentation by iteratively merging neighboring homogeneous regions that had frequently co-occurring region types. In both approaches, certain segments in certain scales may correspond to compound structures, but the grouping criteria still do not involve spatial arrangements and hence may fail in detecting and delineating many other structures.

Another problem with tile-based modeling is the assumption that the whole window corresponds to a compound structure where feature extraction is performed holistically. To identify structure-sensitive neighborhoods, Vanegas *et al.* [12] proposed a graph-based method to determine aligned groups of objects from a given segmentation. However, this method was designed for specific arrangements such as alignment and parallelism. It also worked in a single scale and was sensitive to segmentation errors. The use of multiple partitionings of the image via segmentation hierarchies has been identified as an important problem in remote sensing. However, it is mainly addressed as the problem of selecting individual regions from a set of candidates [13]–[17] with no consideration of the contextual interactions between neighboring regions.

In this paper, we propose a generic method for the modeling and detection of compound structures that can involve the arrangements of an unknown number of primitive objects. The procedure starts with a single example compound structure that contains primitive objects that are used to estimate a probabilistic appearance and arrangement model. The modeling process considers the primitive objects as random variables in a Markov random field (MRF) where potentially related objects are connected. MRFs have been used in the literature to model contextual information in neighborhoods of pixels [18] or regions [19], [20]. Our aim is to learn a flexible arrangement model with a small number of examples that can distinguish between different types of compound structures inside a large scene instead of dedicating the MRF to model the whole scene with only a limited set of relationships. The parameters of the proposed MRF model are learned via sampling from the corresponding maximum entropy distribution.

The detection task is formulated as the selection of multiple coherent subsets of candidate regions obtained from a hierarchical segmentation where each set of selected regions, when grouped together, constitutes an instance of the example compound structure. This differs from our earlier work [3] that did not need an initial segmentation of the primitives but required that their number is given *a priori*. The proposed selection algorithm models the spatial relationships among the candidate regions by using the multiscale neighborhood graph. Our algorithm uses a sampling procedure to maximize the likelihood of groups of regions where the decision of selecting or not selecting regions is done jointly as groups instead of individual decisions. Furthermore, our algorithm does not have any *a priori* knowledge of the number of regions to be selected. It also enables the detection of regions that cannot be detected by using spectral and shape features alone, owing to the contextual information that the model captures. In summary, our major contributions are threefold. First, we describe a model for the individual appearance properties of primitive objects as well as their spatial arrangements within compound structures. Second, we propose a solution to the combinatorial region selection

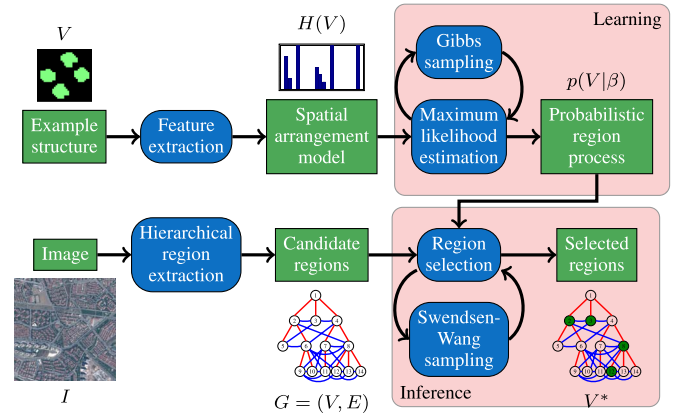


Fig. 2. Object/process diagram of the proposed approach. Rectangles represent objects, and rounded rectangles represent processes. The details of all steps are presented in the following sections.

problem for detecting and localizing an unknown number of instances of a given compound structure in a large scene. Third, to avoid the over- or under-segmentation of candidate regions, we seamlessly integrate multi-scale information and search for the most meaningful regions appearing at different scales of a hierarchical segmentation.

An overview of the proposed approach is shown in Fig. 2. The rest of this paper is organized as follows. Section II introduces the representation for primitive objects and the probabilistic model for their spatial arrangement and shape characteristics. Section III describes the learning algorithm for the estimation of the parameters in the proposed model. Section IV describes the selection algorithm for finding the structures with similar arrangements among a set of candidate regions. Section V presents experimental results, followed by conclusions in Section VI.

II. COMPOUND STRUCTURE MODEL

Compound structures arise from local interactions between primitive objects as well as their individual properties. The set of factors that make the individual primitives members of a compound structure can be motivated by the Gestalt rules that attempt to model the perceptual grouping process in the human vision system. In the following, we present the representation for the primitives, propose a generic spatial arrangement model for grouping these primitives according to semantic cues such as proximity, continuity, parallelism, alignment, etc., and describe a statistical model that encodes the spatial arrangement properties of these groupings into a probabilistic region process.

A. Primitive Representation

In this paper, compound structures are defined as high-level heterogeneous objects that are composed of spatial arrangements of multiple, relatively homogeneous, and compact primitive objects. The set of primitives includes objects that can be relatively easily extracted using low-level operations that exploit spectral, textural, or morphological information. These objects, such as buildings and trees, can be used as building blocks of more complex structures. In this paper, each

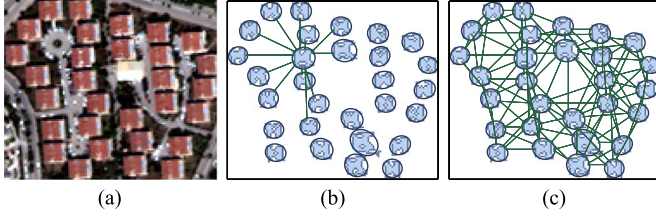


Fig. 3. Neighborhood graph. (a) RGB image. (b) (Blue ellipses) Primitive objects and (green lines) edges representing the neighbors of one primitive. (c) Graph for all primitives.

primitive object v_i is represented by an ellipse $v_i = (l_i, s_i, \theta_i)$, where $l_i = (l_i^x, l_i^y) \in [0, X_{\max} - 1] \times [0, Y_{\max} - 1]$ represents the ellipse's center location, $s_i = (s_i^h, s_i^w) \in [s_{\min}^h, s_{\max}^h] \times [s_{\min}^w, s_{\max}^w]$ contains the ellipse's major and minor axis lengths, respectively, and $\theta_i \in [0, \pi)$ is the orientation measured as the angle between the major axis of the ellipse and the horizontal image axis. Here, X_{\max} and Y_{\max} are the width and height of the image, respectively, and (s_{\min}^h, s_{\max}^h) and (s_{\min}^w, s_{\max}^w) are the minimum and maximum major and minor axis lengths, respectively.

Ellipses have often been used as the image primitives in perceptual organization [21] and object recognition [22] tasks in the computer vision literature, and the underlying assumption that the primitives have relatively compact shapes also holds for many objects of interest in remotely sensed scenes. Ellipses provide simple but sufficiently flexible approximations that can model the most fundamental object characteristics like location, scale, and orientation and can generalize to other shapes such as circles, rectangles, and line segments with additional constraints on specific parameters. The following sections show that they also enable effective and efficient feature extraction and model estimation steps.

B. Spatial Arrangement Model

For a given compound structure consisting of M primitive objects, we construct a neighborhood graph $G = (V, E)$, where the vertices $V = \{v_1, \dots, v_M\}$ correspond to the individual primitive objects and the edges E model their spatial relationships (see Fig. 3). The neighborhood information is obtained by proximity analysis where a threshold on the distance between the closest pixels of each object pair is used to determine the neighbors. In particular, let P_i denote the set of pixels inside the ellipse v_i . Then, $(v_i, v_j) \in E$ if and only if the distance between the closest pixels of v_i and v_j is less than a proximity threshold δ , i.e., $E = \{(v_i, v_j) \in V \times V : \exists (p_i, p_j) \in P_i \times P_j \text{ such that } \forall (p'_i, p'_j) \in P_i \times P_j, d(p_i, p_j) \leq d(p'_i, p'_j) \text{ and } d(p_i, p_j) \leq \delta\}$ where $d(p_i, p_j)$ denotes the Euclidean distance between two pixels p_i and p_j .

For each neighboring primitive object pair $(v_i, v_j) \in E$, we compute the following four features (see Fig. 4):

- 1) distance between the closest pixels, $\phi_{ij}^1 = \min_{p_i \in P_i, p_j \in P_j} d(p_i, p_j)$;
- 2) relative orientation, $\phi_{ij}^2 = \min\{|\theta_i - \theta_j|, 180 - |\theta_i - \theta_j|\}$;
- 3) angle between the line joining the centroids of the two objects and the major axis of a reference object, $\phi_{ij}^3 =$

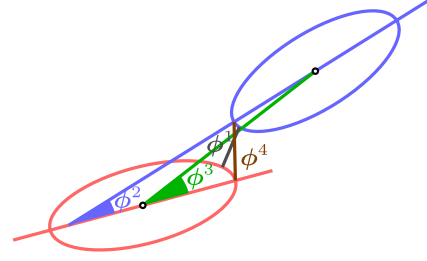


Fig. 4. Pairwise feature examples. $\phi^1, \phi^2, \phi^3, \phi^4$ are described in the text.

$\min\{|\alpha_{ij} - \theta_i|, 180 - |\alpha_{ij} - \theta_i|\}$, where α_{ij} is the angle of the line segment connecting the centroids of v_i and v_j ;

- 4) distance between the closest antipodal pixels that lie on the major axes, $\phi_{ij}^4 = \min_{p_i \in P_i^a, p_j \in P_j^a} d(p_i, p_j)$, where P_i^a denotes the two antipodal pixels on the major axis of v_i .

These features capture various Gestalt properties such as proximity, parallelism, directional continuity, and proximal continuity, respectively. Furthermore, ϕ^2 and ϕ^3 together measure how much the two objects are aligned. In addition to the pairwise features, we also compute the following two individual features for each primitive object v_i :

- 1) area, $\phi_i^5 = \pi(s_i^h/2)(s_i^w/2)$;
- 2) eccentricity, $\phi_i^6 = \sqrt{1 - (s_i^w/s_i^h)^2}$.

Then, given the set of primitives V and the corresponding features, a 1-D marginal histogram H^k is constructed for each $\phi^k, k = 1, \dots, 6$, calculated over all V and E . We append all marginal histograms and use $H(V) = (H^1(E), H^2(E), H^3(E), H^4(E), H^5(V), H^6(V))^T$, where E is assumed to be deterministically computed from V , as a nonparametric approximation to the distribution of the feature values of the primitive objects in the compound structure. The vector length $|H(V)|$ is the total number of bins in all marginal histograms.

C. Probabilistic Region Processes

The diversity of the patterns in different scenes and the richness of the details in each scene entail the use of statistical approaches. In our model, each primitive object v_i (i.e., the ellipse parameters) is considered a vector-valued random variable. Hence, a compound structure is represented by a set of random variables that leads to a region process that follows some true unknown distribution.

When there is incomplete information about a probability distribution, it is desired to use the least informative distribution that makes the fewest number of assumptions. The principle of maximum entropy states that the desired distribution is the one that has the largest possible entropy while still being consistent with the information available in the data [23]. Given N independent and identically distributed observations $\mathcal{V} = \{V^1, \dots, V^N\}$ and their histogram-based representations $H(V^n), n = 1, \dots, N$, as described in the previous section, the information in the training data can be summarized using the empirical expectation

$$\mathbb{E}_{\mathcal{V}}[H(V)] = \frac{1}{N} \sum_{n=1}^N H(V^n). \quad (1)$$

The consistency of the desired model with the evidence in the training data can be enforced by equating the expectation

$$\mathbb{E}_p [H(V)] = \int_V H(V)p(V)dV \quad (2)$$

with respect to the model distribution $p(V)$ to the empirical expectation in (1). Then, given \mathcal{P} as the set of all probability distributions on the random variable V , the maximum entropy distribution is obtained as the solution to the constrained optimization problem

$$p^* = \arg \max_{p \in \mathcal{P}} - \int_V p(V) \log p(V) dV$$

subject to $\mathbb{E}_p [H(V)] = \mathbb{E}_V [H(V)]$. (3)

The region process is governed by the optimal solution p^* , which is also known as the Gibbs distribution, and by the calculus of variations, it takes the form

$$p(V|\beta) = \frac{1}{Z_v} \exp \{ \beta^T H(V) \} \quad (4)$$

where $\beta = (\beta^1, \beta^2, \beta^3, \beta^4, \beta^5, \beta^6)^T$ is the parameter vector controlling each histogram bin and Z_v is the partition function [24]. A region process is equivalent to an MRF according to the following proposition.

Proposition 1: Let G define an MRF. p in (4) satisfies the conditional independence properties of G .

Proof: We show that p can be represented as a product of factors, one per maximal clique in the graph. Note that we can restrict the parameterization to the edges and vertices of the graph, rather than the maximal cliques. Let $p(V|\beta) = (1/Z_v) \prod_{e \in E} \varphi^1(e) \varphi^2(e) \varphi^3(e) \varphi^4(e) \prod_{v \in V} \varphi^5(v) \varphi^6(v)$, where Z_v is the partition function. We define the edge and vertex factors as $\varphi^k(e) = \exp\{(\beta^k)^T H^k(e)\}$, $k=1, 2, 3, 4$, and $\varphi^k(v) = \exp\{(\beta^k)^T H^k(v)\}$, $k=5, 6$, where H^k , $k=1, \dots, 6$, are 1-D marginal histograms computed for the features ϕ^k , $k=1, \dots, 6$. The proof is complete by the Hammersley–Clifford theorem [24]. \square

D. Dynamic Topology of Probabilistic Region Processes

Unlike the traditional MRFs, the neighborhood structure of a region process in our model is not determined *a priori*. The topology of the underlying graph depends on the values of the variables in the process. Assigning a new value to a primitive object (e.g., moving, scaling, or rotating the corresponding ellipse) may change its set of neighbors, i.e., produce new neighbors and remove existing ones. An important observation is that using neighborhood structures based on Voronoi tessellations or k -nearest neighbors may cause changes in the neighborhood relations of other variables whenever a variable is modified. Conversely, determining the neighborhood structure using proximity makes the neighborhood relations between the other variables remain unchanged. Using the aforementioned

property and Proposition 1, we derive the following corollary that helps the estimation procedure in the following section.

Corollary 1: The conditional distribution for each individual variable v_i depends only on its neighbors given a realization of the process $V = \{v_1, \dots, v_M\}$ as

$$\begin{aligned} p(v_i|V \setminus v_i) &= \frac{p(V)}{\sum_{v'_i} p(v'_i \cup V \setminus v_i)} \\ &= \frac{\prod_{c_{v_i} \in C(G)} \varphi(c_{v_i}) \prod_{c_{\setminus v_i} \in C(G)} \varphi(c_{\setminus v_i})}{\sum_{v'_i} \prod_{c_{v'_i} \in C(G')} \varphi(c_{v'_i}) \prod_{c_{\setminus v'_i} \in C(G')} \varphi(c_{\setminus v'_i})} \\ &= p(v_i|nb(v_i)) \end{aligned} \quad (5)$$

where $C(G)$ represents the cliques of graph G , c_{v_i} and $c_{\setminus v_i}$ represent each clique that involves and does not involve v_i , respectively, $nb(v_i)$ denotes the neighbors of v_i , and G' in the denominator represents the graph that is formed for the current value of v'_i .

The equality in (5) follows from the observation that all terms that do not involve v_i cancel out between the numerator and denominator, so only the products of cliques that contain v_i are left. However, if we use Voronoi tessellations or k -nearest neighbors, the cancellations would not occur because the $c_{\setminus v'_i}$ would be different for every assignment of v'_i in the summation.

III. LEARNING

A. Maximum Likelihood Estimation

Suppose that we observe a set of region processes $\mathcal{V} = \{V^1, \dots, V^N\}$ that are assumed to be independent and identically distributed realizations of the same compound structure. These observations can be manually marked on an image or drawn by a human analyst. We can estimate a compound structure model via the maximum likelihood estimation (MLE) of the unknown parameter vector β by maximizing the log-likelihood of the data

$$\ell(\beta|\mathcal{V}) = \sum_{n=1}^N \log p(V^n|\beta). \quad (6)$$

The gradient of the log-likelihood is given by

$$\frac{d\ell(\beta|\mathcal{V})}{d\beta} = \mathbb{E}_p [H(V)] - \frac{1}{N} \sum_{n=1}^N H(V^n). \quad (7)$$

Since the MLE problem is differentiable and jointly concave in the vector β , gradient ascent algorithms are guaranteed to converge to the global optimum. We use the stochastic gradient ascent algorithm where the expectation $\mathbb{E}_p [H(V)]$ in (7) is approximated by a finite sum of histograms of samples $V^{(s)}$, $s=1, \dots, S$, drawn independently from the distribution $p(V|\beta)$, as

$$\hat{\mathbb{E}}_p [H(V)] = \frac{1}{S} \sum_{s=1}^S H(V^{(s)}). \quad (8)$$

The pseudocode for the resulting method is shown in Algorithm 1. In the next section, we describe a Markov chain Monte Carlo (MCMC)-based method for generating each sample $V^{(s)}$ in line 5 of the algorithm.

Algorithm 1 Stochastic gradient ascent for MLE of β .

Input: $\mathcal{V} = \{V^1, \dots, V^N\}$

Output: β

- 1: Initialize weights β randomly
 - 2: $\eta \leftarrow 1$
 - 3: **repeat**
 - 4: **for** $s \leftarrow 1$ to S **do**
 - 5: Sample $V^{(s)} \sim p(V|\beta)$
 - 6: **end for**
 - 7: $\hat{\mathbb{E}}_p[H(V)] \leftarrow (1/S) \sum_{s=1}^S H(V^{(s)})$
 - 8: $\beta \leftarrow \beta + \eta(\hat{\mathbb{E}}_p[H(V)] - (1/N) \sum_{n=1}^N H(V^n))$
 - 9: Decrease step size η by a factor of 0.5
 - 10: **until** log-likelihood in (6) unchanged
-

B. Sampling Region Processes

We use a Gibbs sampler that samples a variable conditioned on the values of all the other variables in the distribution parameterized by β in a particular iteration of the stochastic gradient ascent procedure. Given a joint sample $\tilde{V}^t = \{v_1^t, \dots, v_M^t\}$ of M variables at the t th sampling iteration, the next step involves replacing the value of a particular variable v_i^t by a new value v_i^{t+1} drawn from the full conditional distribution $p(v_i | \tilde{V}^t \setminus v_i^t, \beta)$. We move from v_i^t to v_i^{t+1} by sampling only one ellipse component (i.e., either one of l_i , s_i , or θ_i) at a time. That is, we choose either one of l_i , s_i , or θ_i to be updated at random, with equal probability, and then, a candidate value is randomly generated for that component from a uniform proposal distribution over the object parameter space defined in Section II-A. This corresponds to randomly translating, scaling, or rotating an ellipse at each sampling iteration. The new value of the selected component, together with the old values of the remaining components, produces a candidate sample v_i^* . Since the proposal distribution is symmetric, the acceptance probability [25] of the candidate sample is obtained as

$$\alpha = \min \left(1, \frac{p(v_i^* | \tilde{V}^t \setminus v_i^t, \beta)}{p(v_i^t | \tilde{V}^t \setminus v_i^t, \beta)} \right). \quad (9)$$

If the proposal is accepted, v_i^{t+1} is set to v_i^* ; otherwise, v_i^{t+1} stays the same as v_i^t . All the other variables remain unchanged, i.e., $v_j^{t+1} = v_j^t$ for $j \neq i$ and $j = 1, \dots, M$.

By Corollary 1, to sample a variable, we only need to know the values of its neighbors before and after the proposal. Thus, the acceptance probability reduces to $\alpha = \min(1, (p(v_i^* | nb(v_i^*), \beta) / p(v_i^t | nb(v_i^t), \beta)))$. Since p can be represented as a product of potentials over vertices and edges, it can be further shown that $p(v_i | nb(v_i), \beta) = (1/Z_v) \exp\{\beta^T H(v_i \cup nb(v_i))\}$, and we can write $\alpha = \min(1, (\exp\{\beta^T H(v_i^* \cup nb(v_i^*))\} / \exp\{\beta^T H(v_i^t \cup nb(v_i^t))\}))$. As a result, when evaluating α , we do not need to calculate the normalization constant

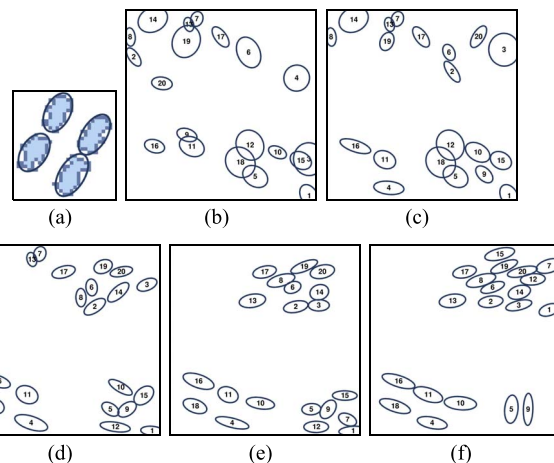


Fig. 5. Illustration of the Gibbs sampler in Algorithm 2. (a) Compound structure V given as input to stochastic gradient ascent in Algorithm 1. (b–f) Samples \tilde{V}^t at iterations $t = 0, 50, 200, 500, 1000$ in Algorithm 2.

Z_v . The sampling procedure is summarized in Algorithm 2 and is illustrated in Fig. 5.

Algorithm 2 Gibbs sampler for producing a particular $V^{(s)}$.

Input: β

Output: $V^{(s)}$

- 1: Initialize $\tilde{V}^0 = \{v_1^0, \dots, v_M^0\}$
 - 2: **for** $t \leftarrow 0, 1, 2, \dots, T-1$ **do**
 - 3: Choose one v_i at random, with equal probability
 - 4: Choose l_i , s_i , or θ_i at random, with equal probability
 - 5: **if** l_i is chosen **then**
 - 6: Sample $l_i^* \sim U([0, X_{\max} - 1] \times [0, Y_{\max} - 1])$
 - 7: $v_i^* \leftarrow (l_i^*, s_i^t, \theta_i^t)$
 - 8: **end if**
 - 9: **if** s_i is chosen **then**
 - 10: Sample $s_i^* \sim U([s_{\min}^h, s_{\max}^h] \times [s_{\min}^w, s_{\max}^w])$
 - 11: $v_i^* \leftarrow (l_i^t, s_i^*, \theta_i^t)$
 - 12: **end if**
 - 13: **if** θ_i is chosen **then**
 - 14: Sample $\theta_i^* \sim U([0, \pi])$
 - 15: $v_i^* \leftarrow (l_i^t, s_i^t, \theta_i^*)$
 - 16: **end if**
 - 17: $v_i^{t+1} \leftarrow \text{UPDATEPRIMITIVE}(v_i^*, \tilde{V}^t, \beta)$
 - 18: $v_j^{t+1} \leftarrow v_j^t$ for $j \neq i$ and $j = 1, \dots, M$
 - 19: **end for**
 - 20: $V^{(s)} \leftarrow \tilde{V}^T$
 - 21: **procedure** UPDATEPRIMITIVE (v_i^*, \tilde{V}, β)
 - 22: Compute $nb(v_i) \in \tilde{V} \setminus v_i$ and $nb(v_i^*) \in \tilde{V} \setminus v_i^*$
 - 23: Compute acceptance probability α
 - 24: Sample $q \sim U(0, 1)$
 - 25: **if** $q < \alpha$ **then**
 - 26: **return** v_i^*
 - 27: **else**
 - 28: **return** v_i
 - 29: **end if**
 - 30: **end procedure**
-

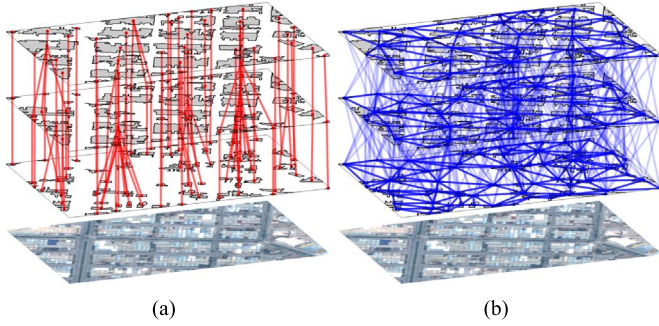


Fig. 6. Hierarchical region extraction. The candidate regions (V) at three levels are shown in gray. (a) Edges that represent parent–child relationship are shown in red. (b) Edges E that represent the final neighbor relationship are shown in blue. For clarity, we do not show the edges between two levels that are not consecutive even though there are edges between all level pairs.

IV. INFERENCE AND REGION SELECTION

Given a compound structure model with learned parameter vector β , we would like to automatically detect all of its instances in an input image I . We first propose a set of candidate primitive regions in the image, and then, an inference algorithm is used to select a coherent subset of those regions that optimize a probability function defined in terms of both appearance and arrangement characteristics of region groups.

A. Hierarchical Region Extraction

The first step involves the identification of primitive regions by using a segmentation algorithm. In this paper, we use opening and closing by reconstruction operations as in [13]. Considering the fact that different objects of interest may appear at different scales, we apply opening and closing by reconstruction using structuring elements in increasing sizes. These operations form a hierarchy in which the regions from all levels are treated as candidate primitives, forming the set $V = \{v_1, \dots, v_M\}$. Fig. 6(a) illustrates the hierarchy.

The next step is to connect the potentially related vertices at all levels to represent the neighbor relationships. Since the candidate regions are fixed at the segmentation step, the set of neighbors for each region can also be fixed, with no need for the dynamic neighborhood definition for the sampling problem in Section III-B. Thus, we use Voronoi tessellations of boundary pixels of regions at each level to identify the neighbors of each region at that level. A Voronoi-based neighborhood definition is preferred at this step as it does not require any parameter like the proximity threshold or the number of neighbors as in the proximity-based and k -nearest neighbor-based definitions, respectively. After computing the Voronoi tessellation at each level of the hierarchy independently, a within-level edge $(v_i, v_j) \in E$ is formed between two vertices if the corresponding regions have neighboring Voronoi cells. Furthermore, a between-level edge $(v'_i, v'_j) \in E$ is also formed if v'_j is at a higher level compared to v'_i and if any descendant of v'_j that is at the same level as v'_i is a Voronoi neighbor of v'_i . Fig. 6(b) illustrates the edges E .

B. Bayesian Formulation

Given a graph $G = (V, E)$ that represents the candidate regions and their neighbor relationships in image I , our goal is to search for coherent groups of regions that attain high probability explanations of instances of compound structures of interest in the image. This problem can be formulated as the selection of a subset V^* among all regions V as

$$V^* = \arg \max_{V' \subseteq V} p(V'|I) = \arg \max_{V' \subseteq V} p(I|V')p(V') \quad (10)$$

where $p(I|V')$ is the observed spectral data likelihood for the compound structure in the image and $p(V')$ acts as the spatial (both shape and arrangement) prior according to the model defined in Section II. We use a simple spectral appearance model where the spectral content of each primitive is assumed to be independent and identically distributed according to a Gaussian with mean μ and covariance Σ so that $p(I|V') = \prod_{v_i \in V'} p(y_i|\mu, \Sigma)$, where y_i is the average spectral vector for the pixels inside the i th region v_i . This formulation assumes that the primitives in a compound structure have similar spectral characteristics as the focus of this paper is to develop a novel spatial data model. Different spectral models will be studied as part of our future work. The spatial appearance probability $p(V')$ is computed as in (4) using ellipses that have the same second moments as the regions in V' .

C. CRF Formulation

The selection problem in (10) can be formulated as a conditional random field (CRF). Let $X = \{x_1, \dots, x_M\}$ where $x_i \in \{0, 1\}$, $i = 1, \dots, M$, be the set of indicator variables associated with the vertices V of G so that $x_i = 1$ implies that region v_i is being selected. Our CRF formulation defines a posterior distribution for hidden random variables X given regions V and their observed spectral features $Y = \{y_1, \dots, y_M\}$ in a factorized form as

$$\begin{aligned} p(X|I, V) &\propto p(I|X, V)p(X, V) \\ &= \frac{1}{Z_x} \prod_{v_i \in V} \exp\{(\psi_i^c + \psi_i^s)x_i\} \prod_{(v_i, v_j) \in E} \exp\{\psi_{ij}^a x_i x_j\} \end{aligned} \quad (11)$$

where the vertex bias terms ψ^c and ψ^s representing color and shape, respectively, and edge weights ψ^a representing arrangement are defined as

$$\psi_i^c = \frac{-1}{2}(y_i - \mu)^T \Sigma^{-1}(y_i - \mu), \quad \forall v_i \in V \quad (12)$$

$$\psi_i^s = \sum_{k=5}^6 \beta_{h^k}^k(\phi_i^k), \quad \forall v_i \in V \quad (13)$$

$$\psi_{ij}^a = \sum_{k=1}^4 \beta_{h^k}^k(\phi_{ij}^k), \quad \forall (v_i, v_j) \in E. \quad (14)$$

The feature ϕ^k is computed via the parameters of the ellipse that has the second moments as the input region, h^k is the index of the histogram bin to which a given feature value belongs in

H^k , and β_j^k denotes the j th component of the parameter vector β^k controlling H^k . Then, selecting V^* in (10) is equivalent to estimating the joint MAP labels given by

$$X^* = \arg \max_X p(X|I, V). \quad (15)$$

D. CRF Inference

The exact inference of (15) is intractable in general graphs, but an approximate solution can be obtained by an MCMC sampler. However, Gibbs sampling that updates one variable at a time can be slow in such models requiring many updates to produce significant changes in the global state, particularly when there is strong dependence between the components [24]. On the contrary, the Swendsen–Wang algorithm [26] mixes much faster by updating the labels of many variables at once.

In this paper, we adapt the Swendsen–Wang algorithm that was designed for the Ising model parameterization, i.e., $\{-1, +1\}$ variables, to sample $\{0, 1\}$ variables. First, the original $\{0, 1\}$ indicator variables X are converted to $\{-1, +1\}$ variables $Z = \{z_i = 2x_i - 1, i = 1, \dots, M\}$. Then, the objective (11) is reformulated by variable substitution as

$$\begin{aligned} p(Z|I, V) &\propto p(I|Z, V)p(Z, V) \\ &= \frac{1}{Z} \prod_{v_i \in V} \exp \left\{ \left(\frac{1}{2}\psi_i^c + \frac{1}{2}\psi_i^s + \frac{1}{4}\psi_i^w \right) z_i \right\} \\ &\quad \prod_{(v_i, v_j) \in E} \exp \left\{ \frac{1}{4}\psi_{ij}^a z_i z_j \right\} \end{aligned} \quad (16)$$

where a new term $\psi_i^w = \sum_{v_j \in V} \psi_{ij}^a$ is added to the vertex biases. We are interested in samples from $p(Z|I, V)$ so that the most likely configuration for Z can be found.

The motivation behind the Swendsen–Wang algorithm is that sampling can sometimes be made easier by adding more variables. Suppose that we introduce auxiliary variables $U = \{u_{ij} : (v_i, v_j) \in E\}$, one per edge, and define the extended model

$$p(Z, U|I, V) \propto p(I|Z, V)p(Z, V)p(U|Z, I, V). \quad (17)$$

A careful selection of $P(U|Z, I, V)$ can make the conditionals $P(U|Z, I, V)$ and $P(Z|U, I, V)$ easy to sample from, and samples for the joint model $P(Z, U|I, V)$ can be obtained by alternately sampling these conditionals with conventional MCMC techniques [27]. Then, marginalization will produce valid Z samples from the original distribution because $\sum_U p(Z, U|I, V) = p(Z|I, V)$.

In the extended model in (17), we assume that u_{ij} are conditionally independent given the vertex variables and are uniformly distributed between 0 and $\exp\{(1/4)\psi_{ij}^a z_i z_j\}$. The conditional distribution of the auxiliary variables can be obtained as

$$\begin{aligned} p(U|Z, I, V) &= \prod_{(v_i, v_j) \in E} \frac{1}{\exp\{\frac{1}{4}\psi_{ij}^a z_i z_j\}} \\ &\quad \mathbb{1} \left[0 \leq u_{ij} \leq \exp\left\{\frac{1}{4}\psi_{ij}^a z_i z_j\right\} \right] \end{aligned} \quad (18)$$

where $\mathbb{1}$ is an indicator function that is 1 when its argument is true and 0 otherwise. Our choice of this $p(U|Z, I, V)$ leads to the joint distribution

$$\begin{aligned} p(Z, U|I, V) &\propto \prod_{v_i \in V} \exp \left\{ \left(\frac{1}{2}\psi_i^c + \frac{1}{2}\psi_i^s + \frac{1}{4}\psi_i^w \right) z_i \right\} \\ &\quad \prod_{(v_i, v_j) \in E} \mathbb{1} \left[0 \leq u_{ij} \leq \exp \left\{ \frac{1}{4}\psi_{ij}^a z_i z_j \right\} \right]. \end{aligned} \quad (19)$$

The conditional distribution of the vertex indicator variables Z given the auxiliary variables U is also obtained as

$$p(Z|U, I, V) \propto p(Z, U|I, V). \quad (20)$$

That is, $p(Z|U, I, V)$ is equal to the product of the selected vertex biases, restricted to the region where all constraints

$$\left\{ 0 \leq u_{ij} \leq \exp \left\{ \frac{1}{4}\psi_{ij}^a z_i z_j \right\}, \quad \forall (v_i, v_j) \in E \right\} \quad (21)$$

are satisfied, and is 0 elsewhere.

In the following, we describe how we sample the extended model via Gibbs sampling from $p(U|Z, I, V)$ and $p(Z|U, I, V)$ alternately. Note that the terms involving the edge weights in (18) can only take two values according to the choice of Z , i.e.,

$$\exp \left\{ \frac{1}{4}\psi_{ij}^a z_i z_j \right\} = \begin{cases} \exp \left\{ \frac{1}{4}\psi_{ij}^a \right\} & \text{if } z_i = z_j \\ \exp \left\{ -\frac{1}{4}\psi_{ij}^a \right\} & \text{if } z_i = -z_j. \end{cases} \quad (22)$$

Consequently, when conditioning on U in (20), the terms $\mathbb{1}[0 \leq u_{ij} \leq \exp\{(1/4)\psi_{ij}^a z_i z_j\}]$ may constrain the allowed combinations of Z . In particular, when $\psi_{ij}^a > 0$:

- if $u_{ij} > \exp\{(-1/4)\psi_{ij}^a\}$, we must have $z_i = z_j$,
- if $u_{ij} \leq \exp\{(-1/4)\psi_{ij}^a\}$, there is no constraint on (z_i, z_j) .

Similarly, when $\psi_{ij}^a < 0$:

- if $u_{ij} > \exp\{(1/4)\psi_{ij}^a\}$, we must have $z_i = -z_j$,
- if $u_{ij} \leq \exp\{(1/4)\psi_{ij}^a\}$, there is no constraint on (z_i, z_j) .

Hence, the selection of U introduces constraints to the distribution giving rise to form connected components of vertices to act as a single bonded unit.

To simplify the notation, we replace each u_{ij} with a binary indicator variable $b_{ij} = \mathbb{1}[u_{ij} > \exp\{(-1/4)\psi_{ij}^a\}]$ that denotes the presence of a bond. The conditional $p(B|Z, I, V)$ for the set of all bond variables $B = \{b_{ij} : (v_i, v_j) \in E\}$ factorizes over the edges as $p(B|Z, I, V) = \prod_{(v_i, v_j) \in E} p(b_{ij}|z_i, z_j, I, v_i, v_j)$. From (22), when $\psi_{ij}^a > 0$

$$\begin{aligned} p(b_{ij} = 1|z_i, z_j, I, v_i, v_j) &= \begin{cases} \frac{\exp\{\frac{1}{4}\psi_{ij}^a\} - \exp\{-\frac{1}{4}\psi_{ij}^a\}}{\exp\{\frac{1}{4}\psi_{ij}^a\}} = 1 - \exp\left\{-\frac{1}{2}\psi_{ij}^a\right\} & \text{if } z_i = z_j \\ 0 & \text{if } z_i = -z_j. \end{cases} \end{aligned} \quad (23)$$

When $\psi_{ij}^a < 0$

$$p(b_{ij} = 1 | z_i, z_j, I, v_i, v_j) = \begin{cases} \frac{\exp\{-\frac{1}{4}\psi_{ij}^a\} - \exp\{\frac{1}{4}\psi_{ij}^a\}}{\exp\{-\frac{1}{4}\psi_{ij}^a\}} = 1 - \exp\{\frac{1}{2}\psi_{ij}^a\} & \text{if } z_i = -z_j \\ 0 & \text{if } z_i = z_j. \end{cases} \quad (24)$$

Sampling from $p(B|Z, I, V)$ and, equivalently, from $p(U|Z, I, V)$ is done by randomly selecting a subset of the bond variables based on $p(b_{ij} | z_i, z_j, I, v_i, v_j)$ and forming sets of connected components \mathcal{C} that are connected by edges with $b_{ij} = 1$. The individual vertices that are not connected to any other vertex are also included in this set. Then, sampling from $p(Z|U, I, V)$ is done by randomly selecting some of these connected components and simultaneously flipping the labels of all vertices within these components so that the constraints

- $z_i = z_j$ if $\psi_{ij}^a > 0$,
- $z_i = -z_j$ if $\psi_{ij}^a < 0$

for $b_{ij} = 1$ are still satisfied. When sampling a connected component $C' \in \mathcal{C}$ from $p(Z|U, I, V)$, the acceptance probability for flipping the labels is given by

$$\gamma(C') = \frac{p(-Z|U, I, C')}{p(-Z|U, I, C') + p(Z|U, I, C')} \quad (25)$$

where

$$p(-Z|U, I, C') = \prod_{v_i \in C'} \exp\left\{\left(\frac{1}{2}\psi_i^c + \frac{1}{2}\psi_i^s + \frac{1}{4}\psi_i^w\right)(-z_i)\right\} \quad (26)$$

is the likelihood of the vertices in C' when their labels are flipped ($z_i \leftarrow -z_i$) and

$$p(Z|U, I, C') = \prod_{v_i \in C'} \exp\left\{\left(\frac{1}{2}\psi_i^c + \frac{1}{2}\psi_i^s + \frac{1}{4}\psi_i^w\right)z_i\right\} \quad (27)$$

is the likelihood when the labels stay the same.

The proposed region selection algorithm is summarized in Algorithm 3 and is illustrated in Fig. 7. We use a simulated annealing procedure [24] as described in Section V to guide the sampling iterations. The sampling procedure continues until the change in the value of the objective (11) between two consecutive iterations is significantly small, and a solution to (15) is obtained by taking the most likely configuration X^* across all samples. Finally, the marginal probabilities for the individual regions in the set V^* that corresponds to this solution are obtained from the frequency of observation of each primitive region during the sampling process.

Algorithm 3 Swendsen–Wang sampler for CRF inference for estimating X^* . The number of iterations R is determined by simulated annealing.

Input: $\psi_i^c, \psi_i^s, \psi_{ij}^a, i, j = 1, \dots, M$

Output: X^*

- 1: Initialize labels $Z = \{z_i = -1, i = 1, \dots, M\}$
- 2: **for** $r \leftarrow 1, 2, \dots, R$ **do**

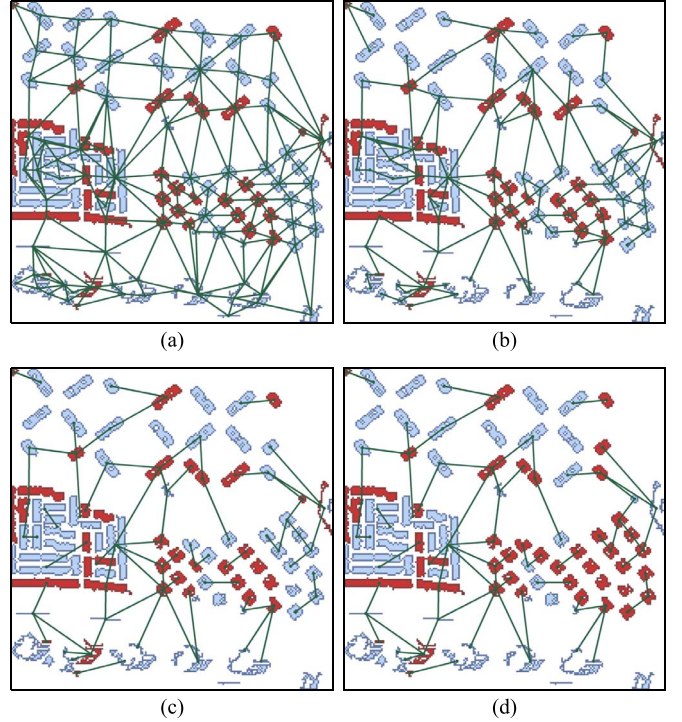


Fig. 7. Illustration of the Swendsen–Wang procedure in Algorithm 3. In each figure, the labels of the primitives are shown in red for selected ($z_i = +1$) and blue for not selected ($z_i = -1$). (a) Labels at the beginning of a particular sampling iteration. The Voronoi edges (E) are shown in green. (b) Edges with positive bond probabilities as candidates for forming connected components of their corresponding vertices. (c) Sampled edges that form connected components of vertices bonded together. (d) Result of randomly flipping the labels of the primitives in some of these components. A single scale is shown for simplicity even though the algorithm normally runs on the graph for the whole candidate region hierarchy.

- 3: **for all** $(v_i, v_j) \in E$ **do**
 - 4: $b_{ij} \leftarrow \text{SAMPLEBONDGIVENVERTICES}(z_i, z_j, \psi_{ij}^a)$
 - 5: **end for**
 - 6: Form connected components \mathcal{C} using bonds $b_{ij} = 1$
 - 7: Pick component $C' \in \mathcal{C}$ uniformly at random
 - 8: Flip labels for all $v_i \in C'$ with probability $\gamma(C')$
 - 9: Compute $X^r = \{x_i = (z_i + 1)/2, i = 1, \dots, M\}$
 - 10: **end for**
 - 11: $X^* \leftarrow \arg \max_{X \in \{X^1, \dots, X^R\}} p(X|I, V)$
 - 12: **procedure** $\text{SAMPLEBONDGIVENVERTICES}(z_i, z_j, \psi_{ij}^a)$
 - 13: **if** $(z_i = z_j \ \& \ \psi_{ij}^a > 0)$ **or** $(z_i = -z_j \ \& \ \psi_{ij}^a < 0)$ **then**
 - 14: Sample $q \sim U(0, 1)$
 - 15: **if** $q < 1 - \exp\{(-1/2)|\psi_{ij}^a|\}$ **then**
 - 16: **return** 1
 - 17: **end if**
 - 18: **end if**
 - 19: **return** 0
 - 20: **end procedure**
-

V. EXPERIMENTS

A. Data Set

The main experiments for quantitative and qualitative evaluation were performed using a multispectral WorldView-2 image of

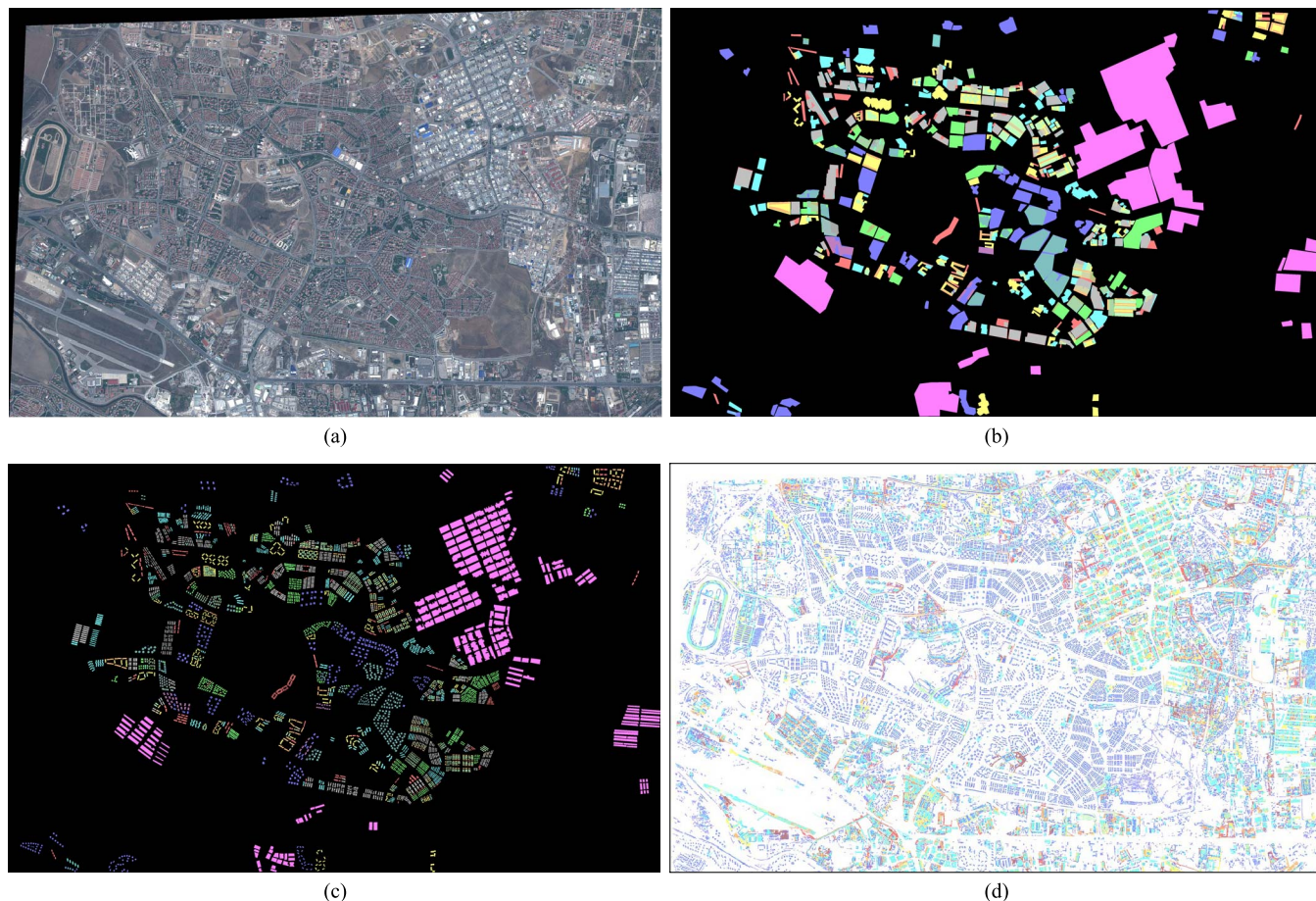








Fig. 8. Data set used for quantitative evaluation. (a) RGB image. (b) Manually delineated polygons reflecting compound structures of interest. (c) Manually delineated buildings inside these polygons. These buildings are used as the primitives in the validation data. The colors of the polygons and buildings correspond to the scenarios given in Table I. (d) Candidate regions obtained by the morphological profile hierarchy. Regions appearing in different levels of the hierarchy are shown with different pseudocolors.

TABLE I
DETECTION SCENARIOS FOR THE EXPERIMENTS. EXAMPLE PRIMITIVES
USED FOR LEARNING THE COMPOUND STRUCTURE MODEL FOR
EACH SCENARIO ARE SHOWN IN A DIFFERENT COLOR.
THE NUMBER OF POLYGONS AND BUILDINGS IN
THE VALIDATION DATA ARE ALSO GIVEN

Scenario	1	2	3	4	5	6
Example primitives						
# polygons	162	98	48	195	60	16
# buildings	1519	870	1117	1796	771	219

Ankara, Turkey. The test scene consisted of 4000×2500 pixels and a 2-m spatial resolution covering various kinds of residential and industrial areas as shown in Fig. 8(a).

The proposed compound structure detection algorithm was evaluated using six scenarios where the first five scenarios correspond to residential structures and the sixth one corresponds to an industrial structure as shown in Table I. All scenarios were formed by various arrangements of four buildings used as the main primitive object of interest in the urban test scene. In particular, the first scenario aimed at the detection of rectangular buildings that are spatially aligned with respect to their major axes. The second scenario aimed at the detection of a structure

composed of buildings placed in a diamond formation. The third scenario aimed at the detection of relatively small dense regularly arranged squarelike buildings. The fourth scenario aimed at the detection of parallel rectangular buildings that are aligned with respect to their minor axes. The fifth scenario aimed at the detection of sparse randomly located squarelike buildings that are slightly larger than those in scenario three. The sixth scenario aimed at the detection of a structure composed of regularly arranged large industrial buildings.

The validation data that were used to evaluate the performance of the method on these scenarios were obtained by the manual delineation of polygons corresponding to compound structures [see Fig. 8(b)] as well as buildings inside these polygons as primitive objects [see Fig. 8(c)]. Table I presents the number of compound structures (polygons) and the corresponding primitives (buildings) in the validation data for each scenario. The learning process for building the compound structure model uses the manual selection of four of these primitives for each structure of interest. This corresponds to triggering the whole learning and inference process using only four individual objects and can be considered a very moderate requirement as only a few individual objects need to be delineated as opposed to relatively large training sets needed for supervised detection and classification algorithms.

B. Experimental Protocol

The experimental procedure for building the example compound structure model (see Section II) and learning its parameters (see Section III) used a single example structure ($N = 1$) with only four primitive objects ($M = 4$) as described earlier. The proximity threshold δ was set to 100 pixels. The corresponding arrangement and shape histograms were constructed with five equal length bins between the minimum and maximum possible values for each feature. The minimum and maximum major and minor axis lengths (s_{\min}^h, s_{\max}^h) and (s_{\min}^w, s_{\max}^w) for sampling the ellipses were both set to (2, 80). This interval was chosen so that it covered the expected smallest and largest primitive axis lengths. The parameters of the maximum entropy model $p(V|\beta)$ were obtained using Algorithm 1. The number of samples S that were used to approximate the expectation $\mathbb{E}_p[H(V)]$ was set to 20. The number of Gibbs sampler iterations T in Algorithm 2 was set to 100.

The experimental procedure for inference and region selection (see Section IV) starts with morphological profiles for hierarchical region extraction. For residential structures, disk structuring elements with radii 2 and 3 were used for constructing the closing profile of the saturation band of the HSV color space computed from the RGB bands of the multispectral image, and for the industrial structures, disk structuring elements with radii from 5 to 10 were used for constructing the opening profile of the HSV value band, as these bands gave good contrast for the primitives of interest (i.e., red roof buildings and industrial buildings, respectively) in our image. A tree structure was constructed from the corresponding profile to extract candidate regions for each scenario. For the residential structures, the number of candidate regions M in two scales was 70 644, and for the industrial structures, the number of candidate regions in six scales was 22 195. This makes a very large pool of candidate regions that we should select from as shown in Fig. 8(d). A Voronoi neighborhood between regions was constructed for each scale, and the neighbors of a region at lower scales were obtained through its descendants in these scales. The resulting graph constructed for the residential scenarios contained 752 754 edges, whereas the graph constructed for the industrial scenario contained 490 222 edges. Considering the total number of the candidate regions in all scales and the number of regions in the validation data, the challenge for the selection problem is that it is expected to select a significantly small fraction of these candidate regions; hence, it should be very selective. Finally, the simulated annealing procedure that was used to help the convergence of Algorithm 3 divided the exponents in the posterior probability in (16) by a certain power called temperature. This temperature was slowly decreased in each iteration according to a cooling schedule such that $\tau_k = 0.995 \tau_{k-1}$ where the initial temperature τ_0 was set to 1.

C. Baselines for Comparison

The first baseline method used sliding windows similar to the tile-based classification tasks in the literature. In particular, we used overlapping 150×150 pixel windows, and using all primitive objects in each window, we extracted marginal histograms, $H(V)$, as described in Section II-B, that modeled

the shape and arrangement characteristics of the primitives at each scale of the hierarchy. Then, we computed the probability that a particular spatial arrangement existed in that window by using $p(V|\beta)$, as described in Section II-C, for each scale and obtained the overall probability for each window as the maximum of the probabilities obtained from all scales. Finally, the marginal probability for each primitive object was obtained as the maximum of the probabilities of the windows that it appeared in. This baseline method aimed to evaluate the effectiveness of the proposed selection process by combining the shape and arrangement information from all primitives.

The second baseline method performed the selection of regions satisfying only color and shape properties by dropping the arrangement terms in the maximum entropy model. Thus, the baseline result was obtained by computing the probability of the candidate regions as $p(X|I, V) \propto (1/Z_x) \prod_{v_i \in V} \exp\{\psi_i^c + \psi_i^s x_i\}$ instead of (11). This choice for the baseline aimed to evaluate the effectiveness of the generic spatial arrangement model in the proposed probabilistic region process compared to the commonly used color and shape-only detectors.

D. Evaluation Criteria

The detection scores resulting from the inference procedure consist of the marginal probabilities of the selected regions (primitives) at the end of Algorithm 3. Thresholding of the score of each region produces a binary detection map. We used precision and recall as the quantitative performance criteria as in [3] and [28] to compare the binary detection maps obtained using a uniformly sampled range of thresholds to the validation data for each scenario that was described in Section V-A. Recall (producer's accuracy), which is computed as the ratio of the number of correctly detected pixels to the number of all pixels in the validation data, can be interpreted as the number of true positives detected by the algorithm, while precision (user's accuracy), which is computed as the ratio of the number of correctly detected pixels to the number of all detected pixels, evaluates the algorithm's tendency for false positives. In addition to the precision–recall curves that used a full range of thresholds, we used a particular threshold value of 0.9 to provide example detection results for all scenarios in the following section. We observed that the particular choice for this threshold was not very critical because, as discussed in the following sections, the inference procedure assigned very high probabilities to most of the selected regions.

Since our selection algorithm detects regions instead of individual pixels, we also performed an object-based evaluation as in [29] in addition to the pixel-based evaluation. This strategy, which is called focus of attention, assumes that a single correctly detected pixel inside a target object is sufficient to attract the operator's attention to that target and label it as correctly detected, but any pixel outside the target is a false alarm because it diverts attention away from true targets. Given the binary detection map for a particular threshold, the union of one or more pixels inside the mask of a validation (ground truth) region was counted as a true positive, while the number of connected components of pixels that did not overlap with any validation region was counted as false positives. Precision

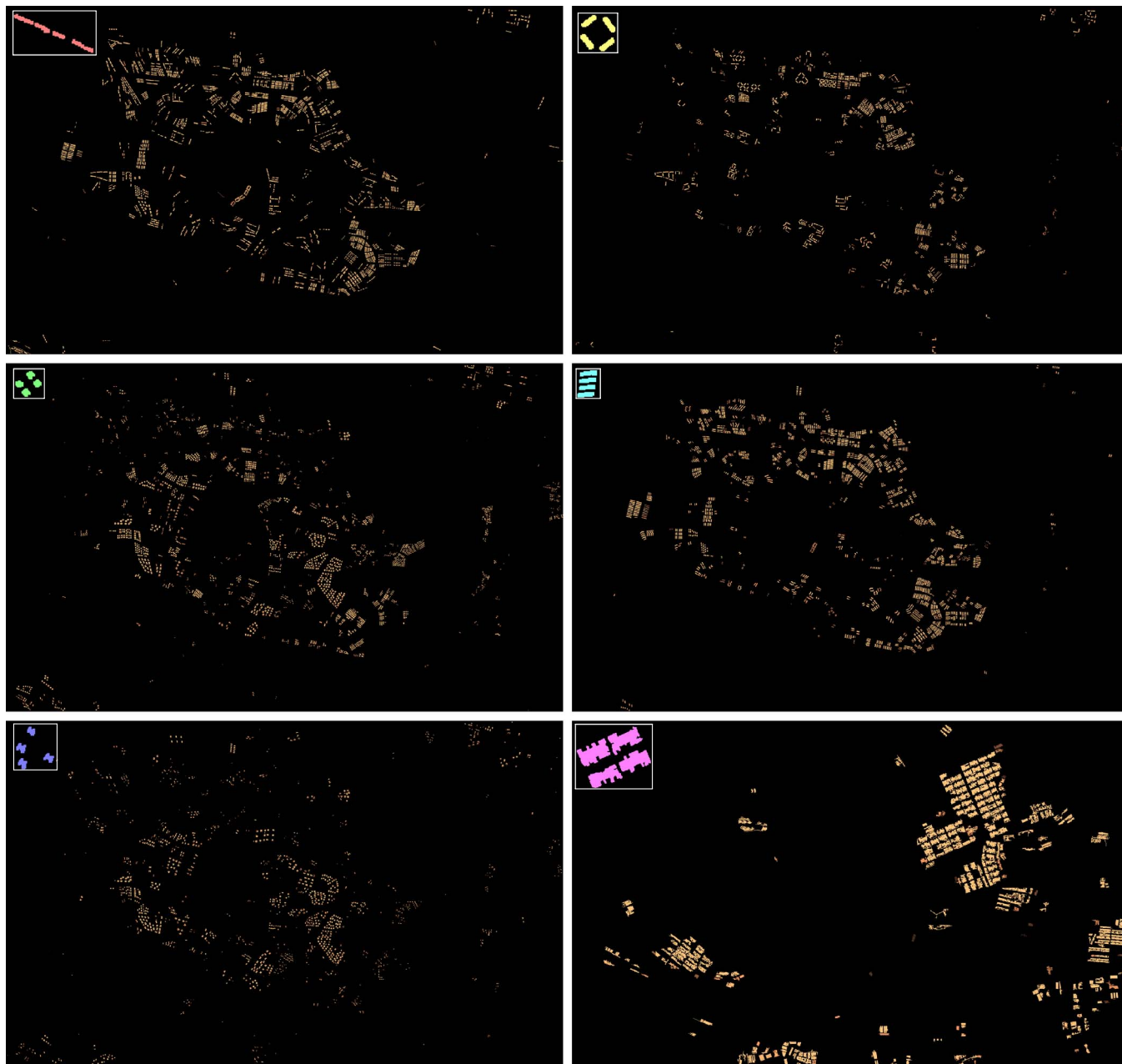


Fig. 9. Marginal probabilities for the selected regions for each scenario. Brighter values indicate higher probabilities. The example primitives are also shown.

and recall used counts of connected groups of pixels instead of individual pixels for object-based evaluation.

E. Results

The learning and inference procedures summarized in Algorithms 1 and 3, respectively, were run for each of the six scenarios on the data set described in Section V-A. The number of selected regions was 3191, 1828, 3819, 3201, 2027, and 1612 for each scenario, respectively. To reconcile the selection of overlapping regions from multiple scales, we computed the maximum of the marginal probability values for each pixel along all scales that it was selected. This operation reduced the number of resulting regions to 1920, 1114, 2648, 1934, 1399, and 357, respectively. These numbers showed that, on the aver-

age, only 4% of all candidate regions in all scales were selected for all scenarios. This meant that most of the regions in the input hierarchy were considered as irrelevant by the proposed method that behaved very selectively even when trained with a single example structure that contained only four buildings for each scenario.

Fig. 9 shows the marginal probabilities of the detected regions for each scenario. The results showed that our selection algorithm was able to detect coherent regions in the image that had arrangements similar to the example structures. Note that a region may belong to more than one type of compound structure as it may form different arrangements with different neighbors. For example, a region may have both close and distant neighbors and may be aligned with different neighbors according to the major and the minor axes at the same time.

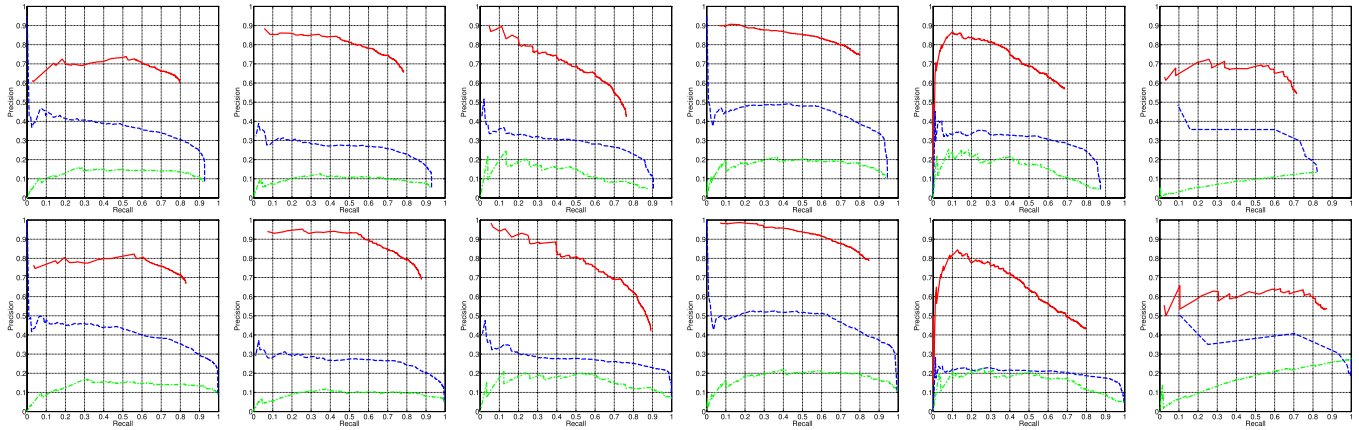


Fig. 10. Precision–recall curves. The columns correspond to scenarios one to six from left to right. The top row corresponds to the pixel-based evaluation, and the bottom row is for the object-based evaluation. The solid red curves correspond to the proposed approach, dashed green ones are for the first baseline (shape and arrangement without selection), and dashed blue ones are for the second baseline (color and shape-only selection with no arrangement).

We observed high marginal probabilities, e.g., greater than 0.9, for most of the selected regions. This indicated that most of the selected regions appeared in most of the sampling iterations, and showed the power of our sampling procedure compared to the traditional Gibbs sampler that samples an individual region at a time by considering only its neighbors. The latter has a potential problem for regions with several irrelevant neighbors that increase the uncertainty in the decision to flip the selection label of a region, whereas our sampling algorithm that sampled connected components and made the decision for a particular region by the contribution of a larger context that contained other regions that might be part of the same structure behaved very selectively. This difference was especially more clear for the boundary regions of compound structures where the marginal probabilities of the boundary regions were as high as the ones in the middle since their decisions were made together through their corresponding connected components.

The next set of experiments was done to compare the performances of the proposed detection algorithm and the baseline methods as described in Section V-B and C, respectively. Fig. 10 shows the precision versus recall curves obtained by applying different thresholds to the marginal probabilities. The results showed that the proposed algorithm that jointly exploited spectral, shape, and arrangement information performed significantly better than the baselines that did not use either selection or arrangement. Even though the two less restricted baselines could approach higher recall levels (bottom right corner of the precision–recall curves) with a sacrifice of substantially reduced precision by accepting more buildings in the output, the proposed method could achieve significantly higher precision values at the same level of recall. The observation that the baseline that used shape and arrangement without selection performed worse than the one that used color and shape-only selection with no arrangement also confirmed the effectiveness of the proposed selection algorithm. When we compared the results for different scenarios, we could observe that the decreases in precision in the third and fifth scenarios were faster than the others for increasing recall (corresponding to decreasing detection threshold). This could be explained by the observation that orientation-based features for squarelike buildings could be

noisy so that more building groups that were not in the validation data appeared in the output as we decreased the detection threshold. This result could also be justified by a smaller ratio of the number of buildings in the validation data versus the number of selections for each of these scenarios.

We also observed that the quantitative evaluation did not always reflect the quality of the results very precisely because the validation data remained approximate. We present zoomed versions of the results for example areas to better illustrate the details for high-resolution imagery. Fig. 11 shows example region hierarchies and selection results. As can be seen in the hierarchies, different regions had better arrangements with their neighbors and had better appearances in different scales with respect to the structure of interest. This fact was reflected in the algorithm by selecting only an appropriate subset of the regions on a path from a leaf region to the highest scale region. Note that misdetections would have occurred if we had manually selected only one scale or attempted to find the single best scale for all the regions. An important property of our algorithm was that it could automatically select regions from different scales. It also did not require *a priori* knowledge of the number of regions to be selected.

Fig. 12 shows more examples of the marginal probabilities and the detections after thresholding these probabilities. The marginal probability values were very strong indicators of the goodness of the detections as the highest likelihood values were obtained for the regions that were very similar to the individual primitives in the example structures and also satisfied the spatial arrangements. On the other hand, the baseline method shown detected a wide range of individual objects without any consideration of their spatial arrangements as expected. This led to very low precision levels as well as unsatisfactory localization of the structures of interest. Furthermore, our method could select regions that would have normally been misdetected if only individual properties were used. For example, structures with diamond formation involved some candidate regions with shorter major axes than the example primitives. The baseline could not detect these regions, whereas our algorithm selected them since their selection along with the others satisfied the arrangement distribution. This was a good example for

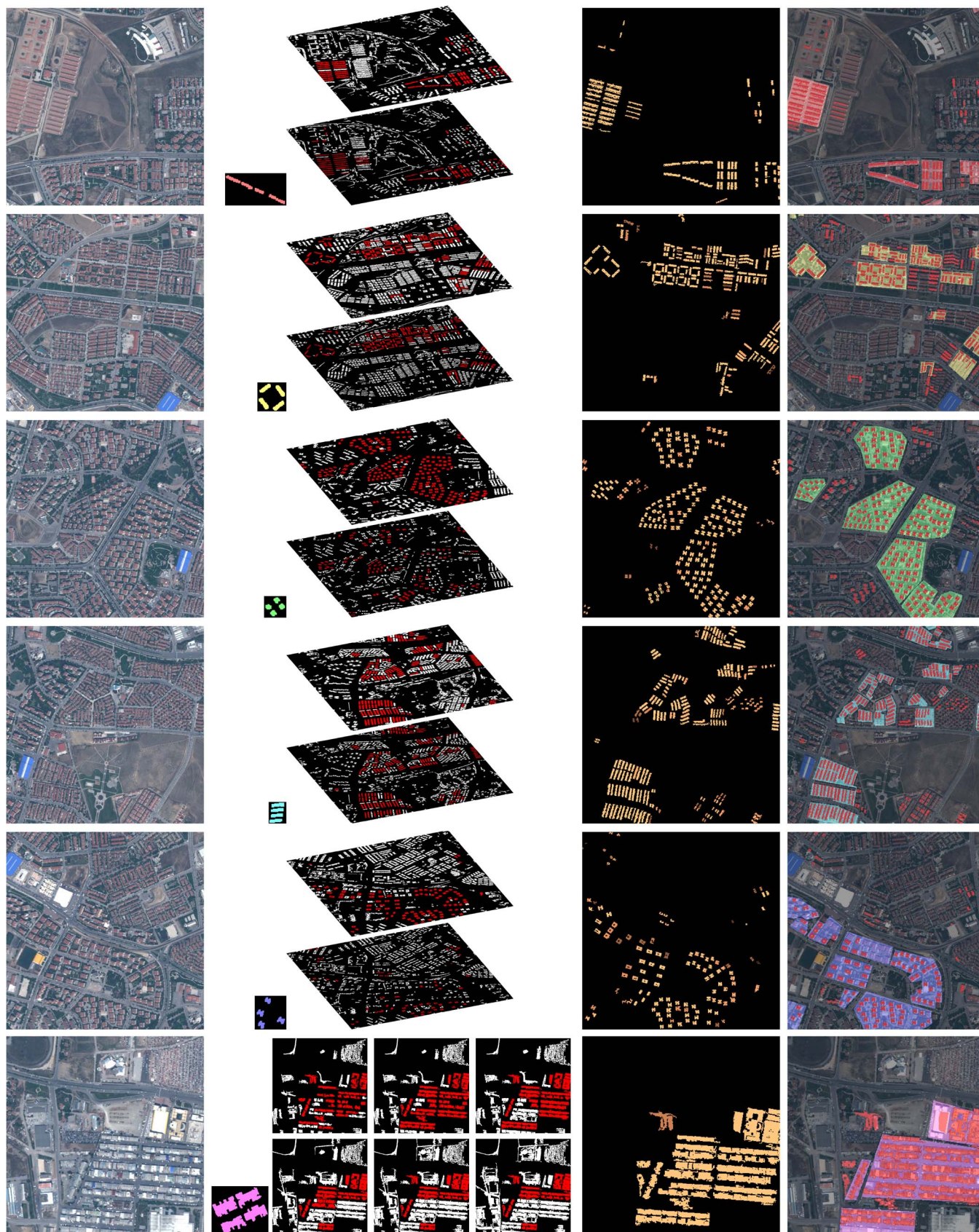


Fig. 11. Zoomed detection examples. The first five rows correspond to the residential structures (scenarios one to five), and the last row corresponds to the industrial structures (scenario six). The first column shows the RGB images for 500×500 subscenes. The second column shows the hierarchy of candidate regions (two-level hierarchy for the first five rows and six-level hierarchy from left to right and top to bottom for the last row). The selected regions are colored with red. The third column shows the marginal probabilities at the end of selection. The fourth column shows the thresholded detections overlaid as red and the validation polygons overlaid with the corresponding colors in Table I.

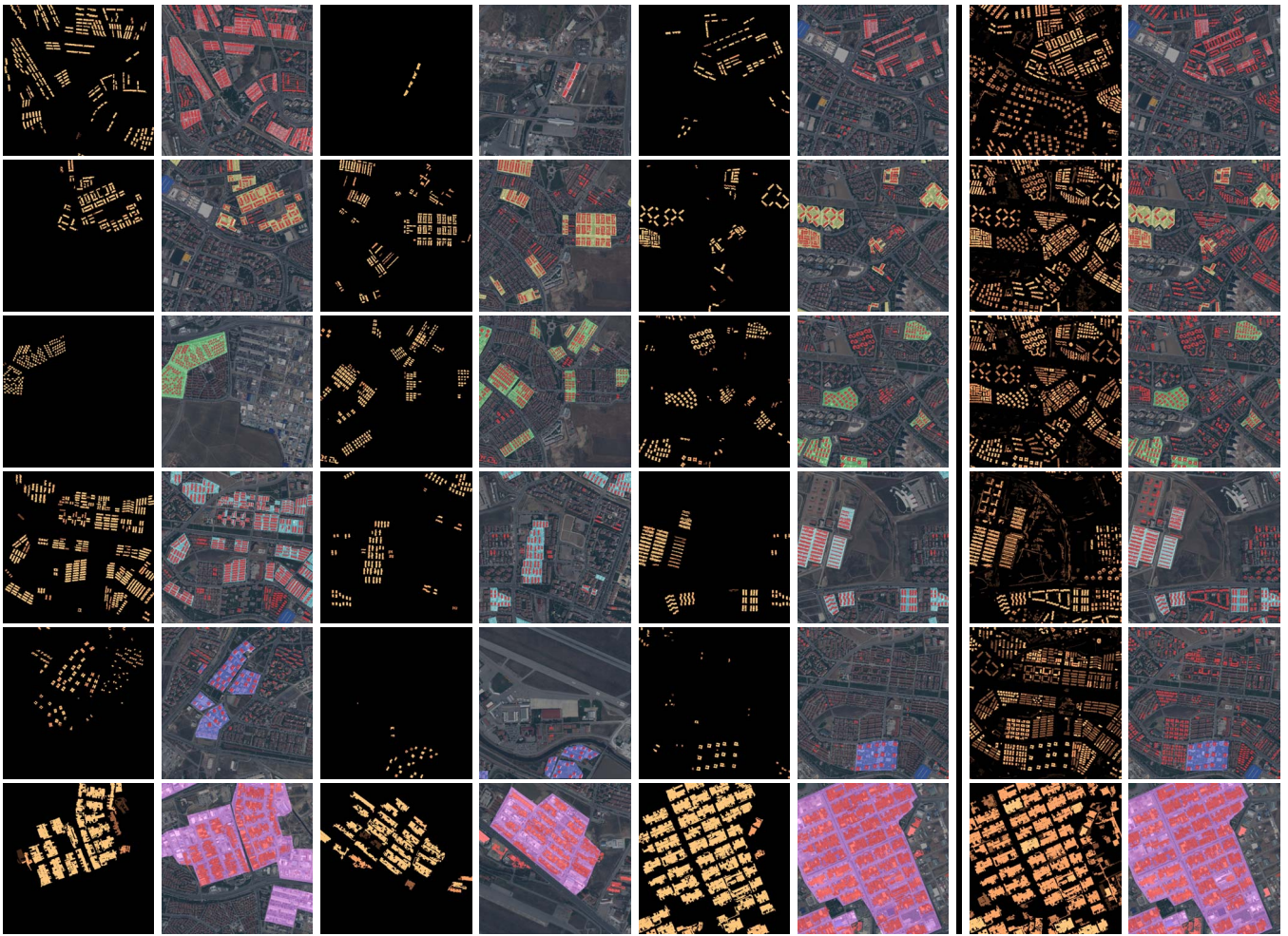


Fig. 12. Additional zoomed detection examples. The image pairs show the marginal probabilities and the overlaid detection results. Each row corresponds to a particular scenario. The first four pairs in each row show the results of our algorithm. The last pair corresponds to the results of the second baseline.

demonstrating the importance of the local spatial context in the selection problem.

We also analyzed different sources of errors in the detections. One of the main reasons for the misdetections was the errors in the input hierarchical segmentation. Some target primitives were never selected because a corresponding candidate region never appeared clearly in the hierarchy. That is, the candidate regions stayed too small until they merged with their surroundings and got completely lost. For example, the industrial regions had complex surfaces that made the morphological operations unable to find some of these regions precisely and prohibited the selection procedure from selecting them. Using additional hierarchical segmentations obtained by different algorithms and/or parameters can overcome this problem by introducing more than one possible set of candidates. Detailed analysis of the results revealed another reason for the misdetections where, even though the arrangements of the candidate regions were satisfying the arrangement distribution of an example scenario, their color and shape properties were not supportive enough for the decision of being selected. Also, in particular, some of the misdetections for the fifth scenario occurred because the primitives were relatively distant from each other. For a candidate region in the image, its closer neighbors might have prevented the

distant neighbors to appear in its Voronoi neighbor set. Then, this region was not selected in the result because it could not connect to the neighbors of interest. Some of the false alarms were caused by single individual regions that had individual statistics that were very similar to those of the example primitives so that the arrangement cues were dominated by the appearance cues. However, since the validation data were subjective, most of the regions that were reflected as false alarms could actually be accepted as true positives under different applications.

In addition to the quantitative experiments using the urban scene in the WorldView-2 image presented in this section, we performed qualitative evaluation by using two additional very high spatial resolution images to illustrate the effectiveness of the proposed approach in detecting different compound structures that are composed of different primitive objects in other types of settings such as agricultural and rural scenes. In particular, we used a multispectral WorldView-2 image of Kusadasi, Turkey, for the detection of fruit orchards as agricultural structures composed of trees as the primitive objects, and we used a panchromatic GeoEye-1 image of Darfur, Sudan, for the detection of refugee camps as rural structures composed of fences as the primitive objects. Example results for orchard detection are presented in Fig. 13. Target orchards are made up of circularly

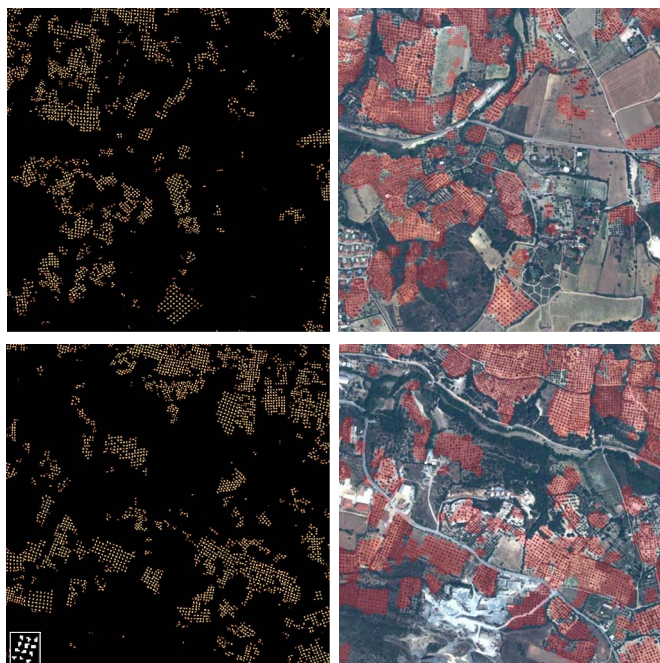


Fig. 13. Example results for the detection of orchards as agricultural structures in two 500×500 pixel WorldView-2 images with 2-m spatial resolution. The left column shows the marginal probabilities at the end of selection. The example primitives used in the learning step are shown on the bottom left corner. The right column shows the thresholded detections overlaid as red. We used a 21×21 pixel Gaussian smoothing filter to enhance the binary detection results before overlaying.

shaped tree primitives appearing in a near-regular repetitive arrangement. Individual trees were localized as candidate regions by using the top-hat transform of the normalized difference vegetation index that had sufficient contrast between the trees and the background. We used a disk structuring element with a radius of 1 pixel in the opening operation. The results show that the method was very successful in identifying the regions corresponding to orchards, with only minor misdetections due to a few missing trees in the top-hat transform outputs.

Example results for the detection of refugee camps are shown in Fig. 14. The goal was to identify the refugee camps consisting of dwellings surrounded by fences made of clay or straw. The fences appear as dark rectangular outlines with one or more entrances (so that the outlines are not closed). More information about the test scene can be obtained from [30]. We aimed to model the fences in terms of spatial arrangements of line segments. Thus, we performed line fitting to the edge detection outputs, and the resulting line segments were considered as candidate primitives in the selection process. The results show that the proposed method could identify the perpendicular arrangement of the fence segments with only a few false positives. A few fence segments could not be detected because they were missing in the line fitting result. Overall, these examples illustrate that the ellipse-based primitive representation and the generic spatial arrangement model together with the proposed learning and inference algorithms were successful in the detection and localization of various compound structures in different types of scenes.

We believe that the output of the proposed method can be particularly useful when the goal is to perform image mining when

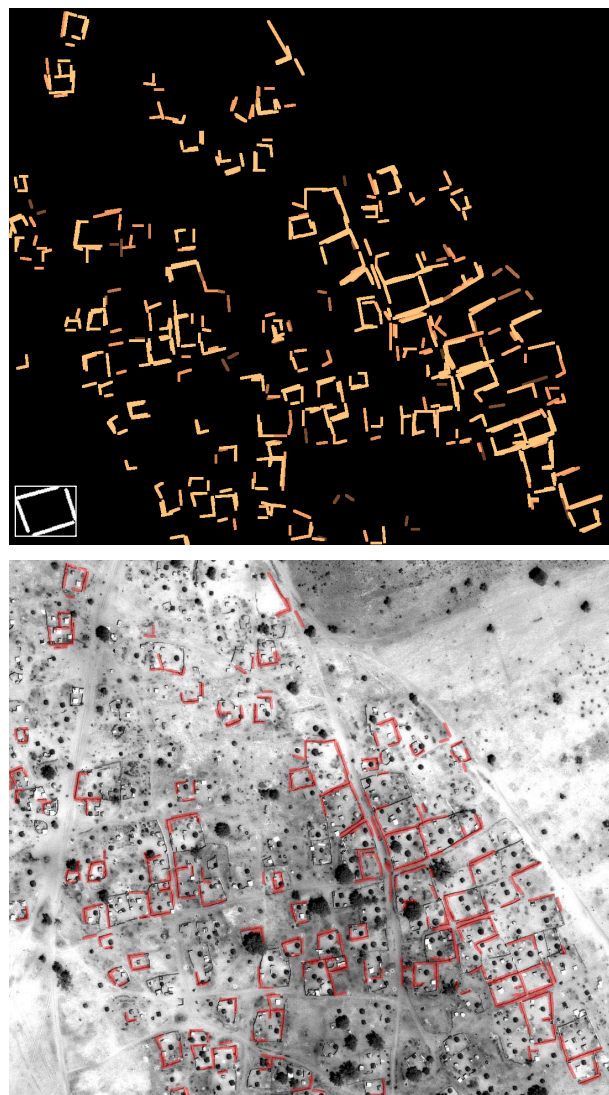


Fig. 14. Example results for the detection of refugee camps as rural structures in a 1102×971 pixel GeoEye-1 image with 0.5-m spatial resolution (GeoEye-1 2009, DigitalGlobe, Inc.). The top image shows the marginal probabilities as well as the example primitives used for learning on the bottom left corner. The bottom image shows the thresholded detections overlaid as red. We used dilation with a disk with radius of 3 pixels to enhance the line segments for display.

we do not have a detailed labeling of example target structures but are interested in finding similar structures using a single example. The localization ability of the algorithm is valuable when there is no clear boundary with respect to low-level cues such as color and texture for the structure of interest. This also conforms to the focus-of-attention strategy that assumes that a single correctly detected pixel inside a target object is sufficient to attract the operator's attention to that target. These results can also be given as input to other algorithms so that more detailed labeling of the image can be produced. For example, the algorithm in [31] aims to estimate the spatial extents of complex geospatial objects that are composed of multiple land use and land cover classes. However, the method requires that at least a single known pixel is given as input for each object so that the procedure can be initialized and the model that was learned from multiple examples can compute its extent. The proposed method can provide the initializations and the models for such

complex structures. As another example, the algorithm in [32] performs detailed classification of urban land use according to the shapes and spatial characteristics of buildings but requires that complete GIS data with individual parcel boundaries and building polygons with detailed attributes are given as input for parcel-based classification where each parcel is assumed to belong to a single homogeneous land use class. The proposed method can localize different urban, rural, or agricultural structures so that the availability of parcel boundaries is no longer a requirement for high-level semantic land use classification.

Finally, we analyzed the execution times for different steps of the proposed algorithm. The proposed learning and inference algorithms were implemented in Matlab with the only exception of the Swendsen–Wang sampling step in Section IV-D implemented in C. We performed a code profile analysis to investigate the time spent in different steps. For the first scenario used in the experiments, the learning process for the example compound structure with four primitives took 774 s using the unoptimized Matlab code on a laptop with a 2.67-GHz Intel Core i5 processor. The sampling process in Section III took 99% of the time where the number of samples was empirically set as described in Section V-B. More samples can take longer but can produce a better model with a higher likelihood. The inference and selection process for an example 500×500 pixel image took 162 s. Of the total time, the hierarchical region extraction in Section IV-A took 1% of the time to produce 854 candidate regions, the feature extraction in Section II-B took 2.5% of the time, finding the Voronoi neighbors in Section IV-A took 26% of the time, and the region selection in Section IV-B took 62% of the time. The hierarchical region extraction step can take longer if the number of scales in the segmentation increases. Using faster algorithms for computing the Voronoi tessellation for the neighborhood graph could decrease the running time.

Another note regarding the implementation is that the proposed algorithm can directly run on large images as the selection algorithm considers only local interactions between the regions within connected components of the large scene graph. However, if the resulting segmentation tree structures are very large with a resulting large number of vertices and edges in the neighborhood graph, sliding windows can be used to process the image where the window size can be selected based on the expected sizes of compound structures. Since the decision in the selection algorithm is based on individual connected components of the graph, it also does not matter how many different structures exist in the same window as the decision for each compound structure is made independently from other structures. This is one of the major advantages over traditional tile-based approaches with tiling being at the core of the image representation where each tile is assumed to correspond to a compound structure and the features are extracted from whole tiles.

VI. CONCLUSION

We described a generic method for the modeling and detection of compound structures that consisted of arrangements of an unknown number of primitive objects in large scenes. The modeling process used a single example structure and built an MRF-based contextual model for the compound structure of

interest whose parameters were learned via sampling from the corresponding maximum entropy distribution. The detection task involved a combinatorial selection problem where multiple subsets of candidate regions from a hierarchical segmentation were selected via the joint sampling of groups of regions by maximizing the likelihood of their individual appearances and relative spatial arrangements. Experiments using very high spatial resolution images showed that the proposed method could effectively localize an unknown number of instances of different compound structures that could not be detected by using spectral and shape features alone.

ACKNOWLEDGMENT

The authors would like to thank Dr. P. Soille from the Institute for the Protection and Security of the Citizen, European Commission Joint Research Centre, Italy, for the Darfur data set.

REFERENCES

- [1] D. Zamalieva, S. Aksoy, and J. C. Tilton, "Finding compound structures in images using image segmentation and graph-based knowledge discovery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2009, vol. 5, pp. 252–255.
- [2] H. G. Akcay and S. Aksoy, "Detection of compound structures using multiple hierarchical segmentations," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2012, pp. 6833–6836.
- [3] C. Ari and S. Aksoy, "Detection of compound structures using a Gaussian mixture model with spectral and spatial constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6627–6638, Oct. 2014.
- [4] R. R. Vatsavai, A. Cheriyyadat, and S. Gleason, "Supervised semantic classification for nuclear proliferation monitoring," in *Proc. IEEE Appl. Imagery Pattern Recog. Workshop*, 2010, pp. 1–10.
- [5] J. Graesser *et al.*, "Image based characterization of formal and informal neighborhoods in an urban landscape," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1164–1176, Aug. 2012.
- [6] L. Gueguen, "Classifying compound structures in satellite images: A compressed representation for fast queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1803–1818, Apr. 2015.
- [7] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.
- [8] B. Demir and L. Bruzzone, "A novel active learning method in relevance feedback for content-based remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2323–2334, May 2015.
- [9] C. Vaduva, I. Gavut, and M. Datcu, "Latent Dirichlet allocation for spatial analysis of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2770–2786, May 2013.
- [10] C. Kurtz, N. Passat, P. Gancarski, and A. Puissant, "Extraction of complex patterns from multiresolution remote sensing images: A hierarchical top-down methodology," *Pattern Recognit.*, vol. 45, no. 2, pp. 685–706, Feb. 2012.
- [11] R. Gaetano, G. Scarpa, and G. Poggi, "Hierarchical texture-based segmentation of multiresolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2129–2141, Jul. 2009.
- [12] M. C. Vanegas, I. Bloch, and J. Inglada, "Alignment and parallelism for the description of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 6, pp. 3542–3557, Jun. 2013.
- [13] H. G. Akcay and S. Aksoy, "Automatic detection of geospatial objects using multiple hierarchical segmentations," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2097–2111, Jul. 2008.
- [14] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.
- [15] Y. Tarabalka, J. C. Tilton, J. A. Benediktsson, and J. Chanussot, "A marker-based approach for the automated selection of a single segmentation from a hierarchical set of image segmentations," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 262–272, Feb. 2012.

- [16] L. Gueguen, G. K. Ouzounis, M. Pesaresi, and P. Soille, "Tree based representations for fast information mining from VHR images," in *Proc. ESA-EUSC-JRC Conf. Image Inf. Mining*, 2012, vol. 1, pp. 15–20.
- [17] B. R. Kiran and J. Serra, "Global-local optimizations by hierarchical cuts and climbing energies," *Pattern Recognit.*, vol. 47, no. 1, pp. 12–24, Jan. 2014.
- [18] G. Moser and S. B. Serpico, "Combining support vector machines and Markov random fields in an integrated framework for contextual image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2734–2752, May 2013.
- [19] J. Porway, Q. Wang, and S. C. Zhu, "A hierarchical and contextual model for aerial image parsing," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 254–283, Jun. 2010.
- [20] 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.
- [21] S. Sarkar and K. L. Boyer, "Perceptual organization in computer vision: A review and a proposal for a classificatory structure," *IEEE Trans. Syst., Man, Cybern.*, vol. 23, no. 2, pp. 382–399, Mar./Apr. 1993.
- [22] A. Y.-S. Chia, D. Rajan, M. K. Leung, and S. Rahardja, "Object recognition by discriminative combinations of line segments, ellipses, and appearance features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1758–1772, Sep. 2012.
- [23] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, Mar. 1996.
- [24] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
- [25] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, Jun. 1953.
- [26] R. H. Swendsen and J.-S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Phys. Rev. Lett.*, vol. 58, no. 2, pp. 86–88, Jan. 1987.
- [27] D. M. Higdon, "Auxiliary variable methods for Markov chain Monte Carlo with applications," *J. Amer. Stat. Assoc.*, vol. 93, no. 442, pp. 585–595, Jun. 1998.
- [28] S. Aksoy, I. Z. Yalniz, and K. Tasdemir, "Automatic detection and segmentation of orchards using very high-resolution imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3117–3131, Aug. 2012.
- [29] N. R. Harvey and J. Theiler, "Focus-of-attention strategies for finding discrete objects in multispectral imagery," in *Proc. SPIE Imaging Spectrom. X*, 2004, vol. 5546, pp. 179–189.
- [30] T. Kemper, M. Jenerowicz, M. Pesaresi, and P. Soille, "Enumeration of dwellings in Darfur camps from GeoEye-1 satellite images using mathematical morphology," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 8–15, Mar. 2011.
- [31] Y. Yang and S. Newsam, "Estimating the spatial extents of geospatial objects using hierarchical models," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2012, pp. 305–312.
- [32] S. Wu, J. Silvan-Cardenas, and L. Wang, "Per-field urban land use classification based on tax parcel boundaries," *Int. J. Remote Sens.*, vol. 28, no. 12, pp. 2777–2800, Jun. 2007.



H. Gökhan Akçay received the B.S. and M.S. degrees in computer engineering from Bilkent University, Ankara, Turkey, in 2004 and 2007, respectively, where he is currently working toward the Ph.D. degree in computer engineering.

His research interests include statistical and structural pattern recognition and computer vision and machine learning for the analysis of medical and remote sensing images.



Selim Aksoy (S'96–M'01–SM'11) received the B.S. degree from the Middle East Technical University, Ankara, Turkey, in 1996 and the M.S. and Ph.D. degrees from the University of Washington, Seattle, WA, USA, in 1998 and 2001, respectively.

He has been working at the Department of Computer Engineering, Bilkent University, Ankara, since 2004, where he is currently an Associate Professor. He spent 2013 as a Visiting Associate Professor at the Department of Computer Science and Engineering, University of Washington. During 2001–2003,

he was a Research Scientist at Insightful Corporation, Seattle, where he was involved in image understanding and data mining research sponsored by the National Aeronautics and Space Administration, the U.S. Army, and the National Institutes of Health. During 1996–2001, he was a Research Assistant at the University of Washington, where he developed algorithms for content-based image retrieval, statistical pattern recognition, object recognition, graph-theoretic clustering, user relevance feedback, and mathematical morphology. During the summers of 1998 and 1999, he was a Visiting Researcher at the Tampere International Center for Signal Processing, Tampere, Finland, collaborating in a content-based multimedia retrieval project. His research interests include computer vision, statistical and structural pattern recognition, machine learning and data mining with applications to remote sensing, medical imaging, and multimedia data analysis.

Dr. Aksoy is a member of the IEEE Geoscience and Remote Sensing Society, the IEEE Computer Society, and the International Association for Pattern Recognition (IAPR). He received the Outstanding Young Scientist Award from the Turkish Academy of Sciences in 2015, the Distinguished Teaching Award from Bilkent University in 2014, a Fulbright Scholarship in 2013, a Marie Curie Fellowship from the European Commission in 2005, the CAREER Award from the Scientific and Technological Research Council of Turkey (TUBITAK) in 2004, and a NATO Science Fellowship in 1996. He was one of the Guest Editors of the special issues on Pattern Recognition in Remote Sensing of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition Letters*, and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2007, 2009, and 2012, respectively. He served as the Vice Chair of the IAPR Technical Committee 7 on Remote Sensing during 2004–2006 and as the Chair of the same committee during 2006–2010. He also served as an Associate Editor of *Pattern Recognition Letters* during 2009–2013.