DETECTION OF COMPOUND STRUCTURES BY REGION GROUP SELECTION FROM HIERARCHICAL SEGMENTATIONS

H. Gökhan Akçay, Selim Aksoy

Department of Computer Engineering
Bilkent University
Bilkent, 06800, Ankara, Turkey
{akcay,saksoy}@cs.bilkent.edu.tr

ABSTRACT

Detection of compound structures that are comprised of different arrangements of simpler primitive objects has been a challenging problem as commonly used bag-of-words models are limited in capturing spatial information. We have developed a generic method that considers the primitive objects as random variables, builds a contextual model of their arrangements using a Markov random field, and detects new instances of compound structures through automatic selection of subsets of candidate regions from a hierarchical segmentation by maximizing the likelihood of their individual appearances and relative spatial arrangements. In this paper, we extend the model to handle different types of primitive objects that come from multiple hierarchical segmentations. Results are shown for the detection of different types of housing estates in a WorldView-2 image.

Index Terms— Contextual modeling, Markov random field, object detection, spatial relationships

1. INTRODUCTION

A challenging problem in remote sensing image information mining is the detection of heterogeneous compound structures such as different types of residential, industrial, and agricultural areas that are comprised of spatial arrangements of simple primitive objects such as buildings and trees. A popular approach for the detection of high-level structures is to divide images into tiles and classify these tiles according to their features. One of such window-based approaches, called the bag-of-words (BoW) model, has been commonly used in recent years for modeling the tile content [1, 2, 3]. However, the BOW representation cannot often effectively model the spatial arrangements which can be the key to detecting many types of compound structures. As an example for exploiting the spatial structure, Vaduva *et al.* [4] modeled relative positions between objects by extracting object pair signatures as words that characterize the tiles. However, the tile-based approaches assume that the whole window corresponds to a compound structure and all of the features inside the window contribute to the modeling of the structure. Consequently, this may result in using many features that are irrelevant to the compound structure of interest. An alternative to tile-based neighborhoods is to use segmentation to identify locally adaptive neighborhoods. Using hierarchical segmentations [5] as multi-scale candidates for meaningful image objects has received significant attention as a potential solution to object detection in remote sensing. However, local spatial arrangements of the neighboring objects have not been considered in these methods.

In [6], we described a generic method for the modeling and detection of compound structures that are comprised of spatial arrangements of an unknown number of primitive objects in very high spatial resolution images. The model considered the primitive objects as random variables, and built a contextual model of their arrangements using a Markov random field. The detection task was formulated as the selection of subsets of candidate regions from a hierarchical segmentation by maximizing the likelihood of their individual appearances and relative spatial arrangements. One limitation of that formulation was that the structures of interest could include only a single type of primitive, e.g., buildings in urban structures. In this paper, we extend our previous work by incorporating additional primitive layers in the modeling and detection process. We show that the use of multiple primitive object layers consisting of multiple hierarchical segmentations provides additional evidence for the detection and localization of the structures of interest, and leads to increased recall compared to simple aggregation of the results where the layers are used independently.

2. COMPOUND STRUCTURE MODEL

The procedure starts with a single example compound structure that contains primitive objects $V = \{v_1, \ldots, v_M\}$ that are used to estimate a probabilistic appearance and arrange-

This work was supported in part by the GEBIP Award from the Turkish Academy of Sciences.

ment model. In particular, we assume that a compound structure V consists of R layers of primitive object maps, $V = \bigcup_{r=1,\ldots,R} V^r$. Each 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. 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.

The modeling process considers the primitive objects (i.e., the ellipses' parameters) V as random variables corresponding to the vertices of a Markov random field (MRF) where potentially related objects are connected using undirected edges $E = \bigcup_{r_1,r_2=1,\ldots,R} E^{r_1 r_2}$ where $E^{r_1 r_2}$ denotes the edges between the vertices at layers r_1 and r_2 (Figure 1). Note that, when $r_1 = r_2$, $E^{r_1 r_2}$ represents the edges between the vertices at the same layer. Let P_i denote the set of pixels inside the ellipse v_i. For each connected primitive object pair $(v_i, v_j) \in E$, we compute the following four features:

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

In addition to the pairwise features, we also compute the following two individual features for each primitive object v_i:

- area, $\phi_i^5 = \pi(s_i^h/2)(s_i^w/2)$,
- 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 one-dimensional marginal histogram $H_k^{r_1 r_2}(E^{r_1 r_2})$ is constructed for each feature $\phi^k, k = 1, \ldots, 4$, computed over all edges for each pair of layers r_1 and r_2. Also, a one-dimensional marginal histogram $H_k^r(V^r)$ is constructed for each feature $\phi^k, k = 5, 6$, computed over all vertices at each layer V^r. The concatenation $H(V)$ of all marginal histograms $H_k^{r_1 r_2}(E^{r_1 r_2}), k = 1, \ldots 4, r_1, r_2 = 1, \ldots, R$, and $H_k^r(V^r), k = 5, 6, r = 1, \ldots, R$, is used as a non-parametric approximation to the distribution of the feature values of the primitive objects in the compound structure. The process is governed by the Gibbs distribution, and takes the form

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

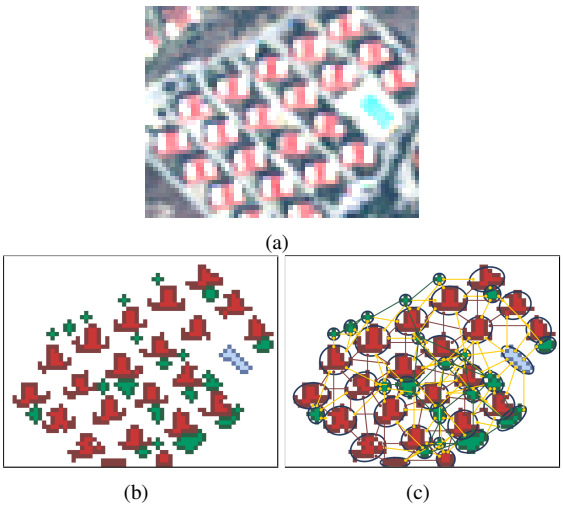
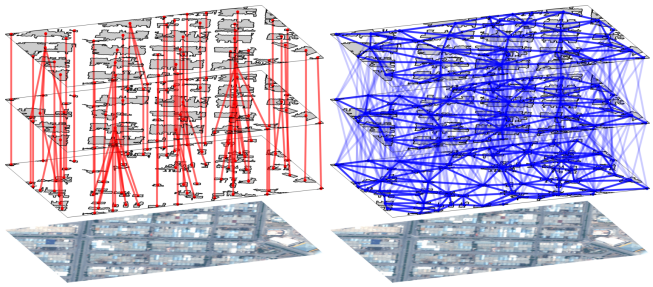


Fig. 1. Neighborhood graph. (a) RGB image. (b) Primitive objects from three different layers: buildings (red), vegetation (green), pool (blue). (c) Graph vertices (blue ellipses) and the edges that connect the primitives in the same layer (red edges for buildings and green edges for vegetation) and between different layers (yellow edges).

where β is the parameter vector controlling each histogram bin, and Z_v is the partition function. The parameters of the proposed MRF model are learned via Gibbs sampling. This corresponds to randomly translating, scaling, or rotating an ellipse at each sampling iteration. Please see [6] for details of the learning algorithm when a single layer is used.

3. DETECTION PROCEDURE

The detection problem is posed as the selection of multiple subgroups of candidate regions $V = \{v_1, \ldots, v_M\}$ coming from multiple hierarchical segmentations where each selected group of regions constitutes an instance of the example compound structure in the large image. The first step in the detection procedure involves the identification of primitive regions for each layer V^r by using a hierarchical segmentation algorithm. The union of these regions from all levels at all layers are treated as candidate primitives, forming the set $V = \bigcup_{r=1,\ldots,R} V^r$. Then, the input hierarchical forest structure is extended by connecting neighboring candidate regions at all levels and all layers with edges E. For each layer, we use Voronoi tessellations of boundary pixels of regions at each level to identify the edges $(v_i, v_j) \in E$ at that level. 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'. For each pair of layers V^{r_1} and V^{r_2}, vertices $v_i^{r_1}$ and $v_j^{r_2}$ are connected with a between-layer edge $(v_i^{r_1}, v_j^{r_2}) \in E$ if the distance between the closest pixels of these objects is less than a proximity threshold. Figure 2 illustrates a hierarchy.



(a) (b)

Fig. 2. Hierarchical region extraction. The candidate regions (V) at three levels of the same layer are shown in gray. (a) The edges that represent parent-child relationship are shown in red. (b) The between-level edges 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 levels (taken from [6]). The extension in this paper involves several of such hierarchies where vertices are connected with edges between spatially close regions.

Given a graph $G = (V, E)$ that represents the candidate regions and their neighbor relationships in image I, the problem can be formulated as the selection of a subset V^* among all regions V as

$$V^* = \arg\max_{V' \subseteq V} p(V'|I) = \arg\max_{V' \subseteq V} p(I|V')p(V') \quad (2)$$

where $p(I|V')$ is the observed spectral data likelihood for the compound structure in the image, and $p(V')$ acts as the spatial prior according to the learned appearance and arrangement model. We use a simple spectral appearance model where the spectral content of each primitive region in a particular layer r is assumed to be independent and identically distributed according to a Gaussian with mean μ_r and covariance Σ_r, so that $p(I|V') = \prod_{r=1,\ldots,R} \prod_{v_i \in V^r} p(y_i|\mu_r, \Sigma_r)$ where y_i is the average spectral vector for the pixels inside the i'th region v_i. The spatial appearance probability $p(V')$ is computed as in (1) using ellipses that have the same second moments as the regions in V'.

We formulate the selection problem in (2) using a conditional random field (CRF). Let $X = \{x_1, \ldots, x_M\}$ where $x_i \in \{0, 1\}, i = 1, \ldots, M$, be the set of indicator variables associated with the vertices V of G so that $x_i = 1$ implies region v_i being selected. Our CRF formulation defines a posterior distribution for hidden random variables X given regions V and their observed spectral features $Y = \{y_1, \ldots, y_M\}$ in a factorized form as

$$\begin{aligned} p(X|I, V) &\propto p(I|X, V)p(X, V) \\ &= \frac{1}{Z_x} \prod_{v_i \in V} \exp\left\{(\psi_i^c + \psi_i^s)x_i\right\} \prod_{(v_i, v_j) \in E} \exp\left\{\psi_{ij}^a x_i x_j\right\} \end{aligned} \quad (3)$$

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

$$\psi_i^c = \frac{-1}{2}(y_i - \mu_r)^T \Sigma_r^{-1}(y_i - \mu_r), \quad \forall v_i \in V^r, \quad (4)$$

$$\psi_i^s = \sum_{k=5}^{6} \beta^r_{k, h_k^r(\phi_i^k)}, \quad \forall v_i \in V^r, \quad (5)$$

$$\psi_{ij}^a = \sum_{k=1}^{4} \beta^{r_1 r_2}_{k, h_k^{r_1 r_2}(\phi_{ij}^k)}, \quad \forall (v_i^{r_1}, v_j^{r_2}) \in E, \quad (6)$$

for $r, r_1, r_2 = 1, \ldots, R$. The feature ϕ^k is computed by using the parameters of the ellipse that has the second moments as the input region, h_k^r is the index of the histogram bin to which a given feature value belongs in H_k^r, and $\beta_{k,j}^r$ denotes the j'th component of the parameter vector β_k^r controlling H_k^r. $h_k^{r_1 r_2}$ and $\beta_{k,j}^{r_1 r_2}$ are defined similarly. Then, selecting V^* in (2) is equivalent to estimating the joint MAP labels given by

$$X^* = \arg\max_X p(X|I, V). \quad (7)$$

Exact inference of the CRF formulation is intractable in general graphs but an approximate solution can be obtained by a Markov chain Monte Carlo sampler. In this paper, we adapt the Swendsen-Wang sampling algorithm that samples the labels of many variables at once. Please see [6] for details of the sampling algorithm when a single layer is used.

4. EXPERIMENTS

We evaluated the proposed approach using a WorldView-2 image of Kusadasi, Turkey. Figures 3 and 4 show two scenarios involving different types of housing estates. The results showed that the earlier version of our algorithm that used only the building layer could not detect several housing estates due to large variations in the spectral appearances of the primitives, but the additional layers such as water and grass gave further evidence for modeling and detecting the compound structures of interest.

5. REFERENCES

[1] J. Graesser, A. Cheriyadat, R. R. Vatsavai, V. Chandola, J. Long, and E. Bright, "Image based characterization of formal and informal neighborhoods in an urban landscape," *IEEE JSTARS*, vol. 5, no. 4, pp. 1164–1176, August 2012.

[2] L. Gueguen, "Classifying compound structures in satellite images: A compressed representation for fast queries," *IEEE TGARS*, vol. 53, no. 4, pp. 1803–1818, April 2015.

[3] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE TGARS*, vol. 51, no. 2, pp. 818–832, February 2013.

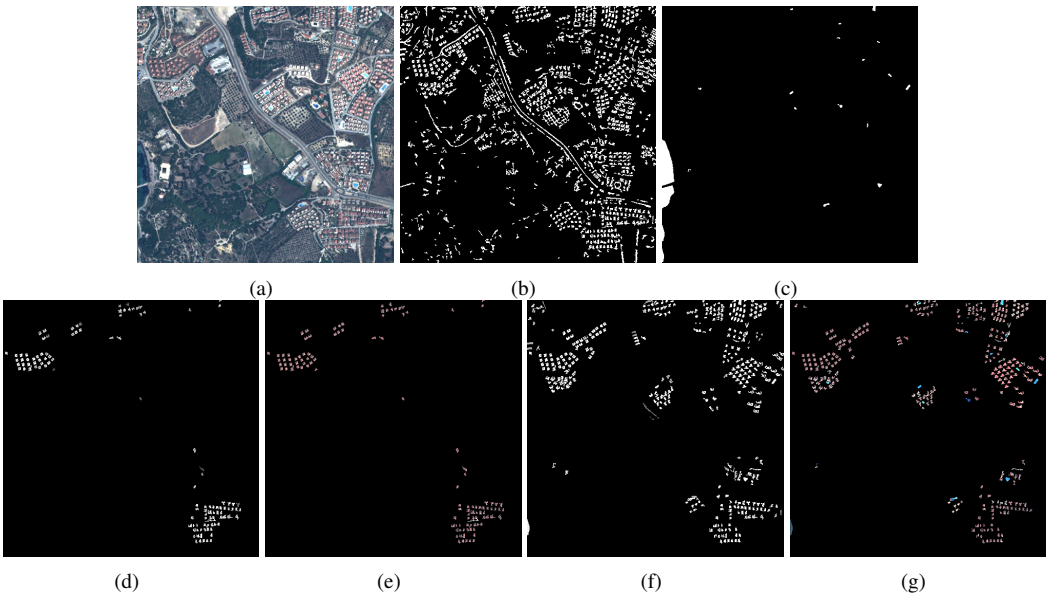


Fig. 3. Example results for detecting housing estates with pools in a 500×500 pixel scene. (a) RGB image. (b) Primitives in the building layer. (c) Primitives in the water layer. (d) Marginal probabilities of the selected regions when only the building layer was used. (e) Masked detections in the RGB image. (f) Marginal probabilities of the selected regions when both the building and the water layers were used. (g) Masked detections in the RGB image.

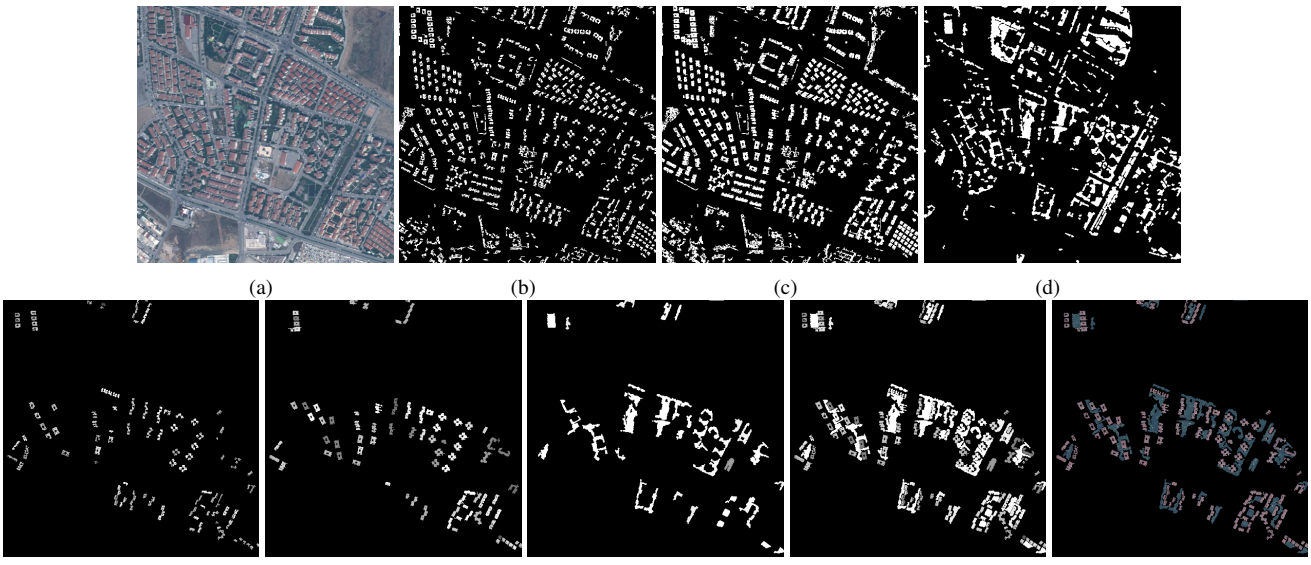


Fig. 4. Example results for detecting housing estates with grass areas in a 500×500 pixel scene. (a) RGB image. (b,c) Primitives in the first and second levels of the building layer, respectively. (d) Primitives in the grass layer. Marginal probabilities of the selected regions when (e) only the first level of the building layer, (f) only the second level of the building layer, (g) only the grass layer, (h) both layers were used. (i) Masked detections in the RGB image.

[4] C. Vaduva, I. Gavat, and M. Datcu, "Latent Dirichlet allocation for spatial analysis of satellite images," *IEEE TGARS*, vol. 51, no. 5, pp. 2770–2786, May 2013.

[5] H. G. Akcay and S. Aksoy, "Automatic detection of geospatial objects using multiple hierarchical segmentations," *IEEE TGARS*, vol. 46, no. 7, pp. 2097–2111, July 2008.

[6] H. G. Akcay and S. Aksoy, "Automatic detection of compound structures by joint selection of region groups from a hierarchical segmentation," *IEEE TGARS*, vol. 54, no. 6, pp. 3485–3501, June 2016.