Detection of Compound Structures Using a Gaussian Mixture Model with Spectral and Spatial Constraints

Çağlar Arı and Selim Aksoy, *Senior Member, IEEE*

***Abstract*—Increasing spectral and spatial resolution of new generation remotely sensed images necessitate the joint use of both types of information for detection and classification tasks. This paper describes a new approach for the detection of heterogeneous compound structures such as different types of residential, agricultural, commercial, and industrial areas that are comprised of spatial arrangements of primitive objects such as buildings, roads, and trees. The proposed approach uses Gaussian mixture models (GMM) in which the individual Gaussian components model the spectral and shape characteristics of the individual primitives and an associated layout model is used to model their spatial arrangements. We propose a novel expectation-maximization (EM) algorithm that solves the detection problem using constrained optimization. The input is an example structure of interest that is used to estimate a reference GMM and construct spectral and spatial constraints. Then, the EM algorithm fits a new GMM to the target image data so that the pixels with high likelihoods of being similar to the Gaussian object models while satisfying the spatial layout constraints are identified without any requirement for region segmentation. Experiments using WorldView-2 images show that the proposed method can detect high-level structures that cannot be modeled using traditional techniques.**

***Index Terms*—Object detection, Gaussian mixture model, expectation-maximization, constrained optimization, spectral-spatial classification, context modeling**

I. Introduction

Recent increase in both the spatial and the spectral resolution of the images acquired from new generation satellites have enabled new applications in which the increased spatial resolution brings out objects' details that were not previously visible and the increased spectral resolution improves the capability to discriminate the physical characteristics of these details. Consequently, these advances have necessitated new approaches that effectively exploit both the spectral and the spatial information for the detection and classification of objects in these images [1].

A popular approach for joint use of spectral and spatial information in the remote sensing literature is to partition the images into regions and use the spectral properties of the pixels inside these regions for classification [2], [3], [4]. However, the methods that aim to obtain a smooth classification map based on homogeneous regions cannot be easily applied to the detection of a wide range of complex objects such as housing estates, schools, airports, agricultural fields, power plants, and industrial facilities that have more heterogeneous structures.

An expected component in the approaches that strive to detect such complex objects, called *compound structures* in this paper, is a framework that models the spatial arrangements of simpler primitive objects [5]. Approaches for the detection of specific structures such as airports [6] and orchards [7] are available. However, these methods are not generalizable because they heavily rely on the peculiarities of the objects of interest. As a more generic method, Bhagavathy and Manjunath [8] proposed a window-based detector using histograms of Gabor texture features, and applied it to the detection of golf courses and harbors. However, histogram-based methods often cannot effectively capture the spatial structure. Harvey *et al.* [9] developed a facility detection framework where the outputs of pixel-based classifiers were used as auxiliary features that were input as additional data bands to a final pixel-based classifier. They applied this framework to the detection of high schools using athletic fields, parking lots, large buildings, and residential areas as auxiliary features. However, this framework does not explicitly model the spatial arrangements, and need training data for both the auxiliary and the final detectors. Vatsavai *et al.* [10] used a latent topic learning algorithm with spectral, textural, and structural features to categorize image tiles as nuclear plants, coal power plants, and airports. However, the tile-based global features often cannot effectively model the geometries or the spatial relationships of the objects that comprise the complex structures.

As an example for more explicit modeling of the spatial structure, Gaetano *et al.* [11] performed hierarchical texture segmentation by iteratively merging neighboring regions that have frequently co-occurring region types. Zamalieva *et al.* [12] used probabilities estimated from region co-occurrences to construct the edges of a region adjacency graph, and employed a graph mining algorithm to find subgraphs that may correspond to compound structures. Akcay and Aksoy [13] combined spectral and shape characteristics of primitive objects with spatial alignments of neighboring object groups in assigning weights to the edges of a region adjacency graph, and then used graph clustering to identify compound objects. However, all of these approaches require an initial segmentation for the identification of the primitive regions, but accurate segmentation of very high spatial resolution remotely sensed images is still a very difficult problem.

Manuscript received July 30, 2013; revised November 1, 2013. This work was supported in part by the TUBITAK Grant 109E193. S. Aksoy was also supported in part by a Fulbright Visiting Scholar grant.

Ç. Arı is with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, 06800, Turkey. Email: cari@ee.bilkent.edu.tr.

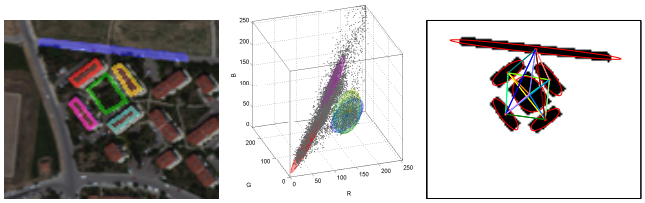
S. Aksoy is with the Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey. Email: saksoy@cs.bilkent.edu.tr.

Using spatial arrangements of local image primitives has also been proven effective for object recognition in the computer vision literature [14]. State-of-the-art methods represent object classes in terms of parts that are visually similar and occur in similar spatial configurations. The detection of parts heavily rely on distinctive invariant features, and their grouping is typically handled using combinatorial searching over all possible part labelings. Popular spatial configurations include the constellation model [15] that is a fully connected joint representation of all parts, and the star model [16] that uses a central reference part and assumes that each part is independent of the others given this part. However, these models are best suited for objects captured from a consistent viewpoint (e.g., sideways cars, frontal faces) in a large amount of training examples, and are not easily applicable to remotely sensed data that contain thousands of primitive objects that do not necessarily generate distinctive features and appear in many different compositions in the overhead view.

This paper describes a new approach that combines spectral and spatial characteristics of simple primitive objects to discover complex compound structures in very high resolution images. The proposed approach uses a probabilistic representation of the image content via Gaussian mixture models (GMM) in which each pixel is represented with a feature vector that encodes both spectral and spatial information consisting of the pixel's multispectral data and its coordinates, respectively. Each Gaussian component in the GMM models a group of pixels corresponding to a particular primitive object where the spectral mean corresponds to the color of the object, the spectral covariance corresponds to the homogeneity of the color content, the spatial mean corresponds to the position of the object, and the spatial covariance models its shape.

The detection procedure starts with a single example compound structure that typically contains a small number of pixels that are used to estimate a reference GMM. This GMM is used to define spectral and spatial constraints for identifying the occurrences of similar compound structures in target images. The spectral constraints ensure that the spectral properties of the detected primitives are similar to those in the reference model, while the spatial constraints assure that the shapes of the detected primitives as well as their spatial layout defined in terms of relative positions are consistent with the reference. We formulate the detection task as a constrained optimization problem that is solved using a novel expectation-maximization (EM) based algorithm that fits a new GMM to the target image data and selects groups of pixels that have high likelihoods of belonging to the Gaussian object models while satisfying the spatial layout constraints. The proposed approach has an important feature that it can localize target structures without any requirement of an initial segmentation. Furthermore, the pixel-based likelihoods computed via the joint use of spectral and spatial information can also handle partial detections and missing primitives, thanks to the contextual information that the model captures. An early version of this paper was presented in [17].

The rest of the paper is organized as follows. Section II defines the proposed compound structure representation. Section III presents the constrained Gaussian mixture model that is used to model this representation. Section IV describes the EM-based maximum likelihood detection algorithm for finding similar structures in target images. Section V presents experimental results using multispectral WorldView-2 imagery. Finally, Section VI provides our conclusions.



(a) (b) (c)

Fig. 1. Illustrations of the spectral and spatial parts of an example compound structure model containing four buildings with a grass area in the middle and a road nearby. (a) Primitive objects overlayed on the RGB image. (b) Spectral model where the gray dots are the pixels' RGB values and the ellipses show the spectral parts of the Gaussians corresponding to the primitive objects. (c) Spatial model where the ellipses overlayed on the binary object masks show the spatial parts of the Gaussians corresponding to the primitive objects and the lines represent the layout model. Note that each primitive object has a corresponding Gaussian in the full spectral-spatial feature space.

II. Definition of Compound Structures

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. To build the model for these structures, first, each pixel is represented by using a d-dimensional feature vector $\mathbf{x} = (\mathbf{x}^{ms}; \mathbf{x}^{xy})$ where $\mathbf{x} \in \mathbb{R}^d$ is formed by concatenating a $d-2$ dimensional vector $\mathbf{x}^{ms}$ containing the multispectral values and a 2-dimensional vector $\mathbf{x}^{xy}$ containing the pixel's coordinates in the image. Since each primitive object is assumed to have a relatively homogeneous spectral content and a compact shape, we further assume that it can be modeled using a Gaussian that is defined in terms of the mean $\boldsymbol{\mu} = (\boldsymbol{\mu}^{ms}; \boldsymbol{\mu}^{xy})$ and the block diagonal covariance matrix $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^{ms}, 0; 0, \boldsymbol{\Sigma}^{xy})$. The covariance model assumes that the multispectral values and the pixel coordinates are independent, i.e., $p(\mathbf{x}) = p(\mathbf{x}^{ms})p(\mathbf{x}^{xy})$, which is similar to the common assumption about the independence of appearance and geometry in the state-of-the-art part-based object detectors [14]. Given a group of pixels forming the primitive object, the spectral mean $\boldsymbol{\mu}^{ms}$ corresponds to the average color of the object, the spectral covariance $\boldsymbol{\Sigma}^{ms}$ corresponds to the homogeneity of the color content, the spatial mean $\boldsymbol{\mu}^{xy}$ corresponds to the position of the object, and the spatial covariance $\boldsymbol{\Sigma}^{xy}$ models its shape as a convex object. Figure 1 illustrates the spectral and spatial parts of an example model.

A compound structure consisting of K primitive objects can then be modeled using a Gaussian mixture model (GMM)

$$p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \alpha_k p_k(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

that is fully defined by the set of parameters $\boldsymbol{\Theta} = \{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ denotes the mean vector and $\boldsymbol{\Sigma}_k \in \mathbb{S}_{++}^d$ denotes the covariance matrix of the k'th

Gaussian component that corresponds to the k'th primitive object. $\alpha_k \in [0,1]$ denotes the probability of a pixel belonging to the k'th component, and is proportional to the number of pixels, i.e., size, of the corresponding primitive object. The sizes are normalized for the whole compound structure, i.e., $\alpha_1, \ldots, \alpha_K$ are constrained to sum up to 1. Since each pixel can belong to one of the K Gaussian components, we also define a corresponding label variable $y_j \in \{1, \ldots, K\}$ for each pixel $j = 1, \ldots, N$ where $y_j = k$ denotes the event of the j'th pixel belonging to the k'th Gaussian component.

The primitive objects can form different compound structures according to different spatial arrangements. In addition to its effectiveness of modeling both the homogeneity and the uncertainty in the spectral and shape content of the primitive objects, the power of the proposed compound structure model comes from its capability of modeling their arrangements. We use a fully connected layout model that is defined in terms of the displacement vectors between the centroids (spatial means) $\boldsymbol{\mu}^{xy}$ of the primitive objects. Given K primitive objects, the spatial layout of the compound structure is modeled using a total of $K(K-1)/2$ displacement vectors $\mathbf{d}_{ij}, i = 1, \ldots, K-1, j = i+1, \ldots, K$, where each of these vectors is defined for a particular pair of primitive objects. Figure 1(c) shows the layout model of the proposed spatial GMM structure.

III. Constrained Gaussian mixture model

In the compound object detection problem, we assume that we are given an example compound structure of interest. This input, called the reference structure, is expected to be in the form of individually delineated regions for the primitive objects. The regions corresponding to the primitive objects can be obtained using basic low-level operations such as morphological opening/closing or image segmentation, or can be obtained via manual selection.

The total of $\tilde{N}$ pixels, $\mathbf{x}_j, j = 1, \ldots, \tilde{N}$, belonging to the reference structure consisting of K primitive objects are used to fit a GMM with K components where each primitive object is modeled by one of the Gaussian components. Since the memberships of all reference pixels to the Gaussian components, $y_j, j = 1, \ldots, \tilde{N}$, are known by definition, the reference GMM parameters can be directly obtained using the maximum likelihood estimates

$$\tilde{\alpha}_k = \frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{N}} [y_j = k] \tag{2}$$

$$\tilde{\boldsymbol{\mu}}_k = \frac{\sum_{j=1}^{\tilde{N}} [y_j = k] \mathbf{x}_j}{\sum_{j=1}^{\tilde{N}} [y_j = k]} \tag{3}$$

$$\tilde{\boldsymbol{\Sigma}}_k = \frac{\sum_{j=1}^{\tilde{N}} [y_j = k] \mathbf{x}_j \mathbf{x}_j^T}{\sum_{j=1}^{\tilde{N}} [y_j = k]} - \tilde{\boldsymbol{\mu}}_k \tilde{\boldsymbol{\mu}}_k^T \tag{4}$$

where $[y_j = k]$ is the Iverson bracket notation whose value is 1 if $y_j = k$, and 0 otherwise. The resulting reference GMM, $p(\mathbf{x}|\tilde{\boldsymbol{\Theta}})$, is defined by its parameters $\tilde{\boldsymbol{\Theta}} = \{\tilde{\alpha}_k, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k\}_{k=1}^K$.

In addition to the GMM parameters, we also extract the spatial layout of the reference structure in terms of the displacement vectors $\tilde{\mathbf{d}}_{ij}, i = 1, \ldots, K-1, j = i+1, \ldots, K$, that are computed using

$$\tilde{\boldsymbol{\mu}}_i^{xy} + \tilde{\mathbf{d}}_{ij} = \tilde{\boldsymbol{\mu}}_j^{xy}. \tag{5}$$

Given a target image with N pixels $\mathbf{x}_j, j = 1, \ldots, N$, the goal is to identify the pixels that are the most similar to those in the reference structure. This can be formulated as a detection problem for the localization of the subregions, i.e., the pixels of interest, that are most likely to correspond to the reference compound object. However, an inherent difficulty in this detection problem is that the pixels of interest, whose number is expected to be similar to the number of pixels in the reference structure, are typically observed as part of a significantly larger set of observations ($N \gg \tilde{N}$) where the rest of the pixels have an unknown distribution. Instead of assuming an explicit density for the whole target image, we define the empirical distribution [18] of the pixels of interest as

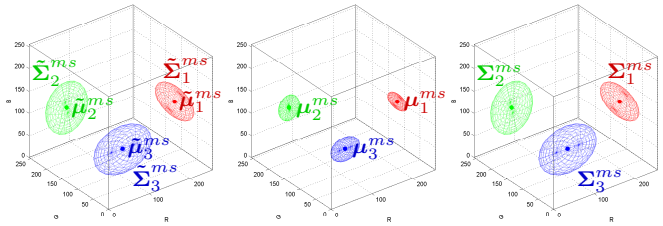
$$\tilde{p}(\mathbf{x}) = \frac{1}{\tilde{N}} \sum_{j=1}^{N} z_j \delta(\mathbf{x} - \mathbf{x}_j) \tag{6}$$

where δ is the Dirac delta function and $z_j \in \{0,1\}, j = 1, \ldots, N$, are the binary indicator variables that identify the pixels of interest, satisfying the constraint $\sum_{j=1}^{N} z_j = \tilde{N}$. $\tilde{p}(\mathbf{x})$ in (6) assigns an equal probability of $1/\tilde{N}$ to the $\tilde{N}$ pixels of interest whose corresponding binary indicator variables z_j are 1, and a probability of 0 is assigned to the remaining points.

The detection process involves modeling the pixels of the target image using a GMM with K components where K is the same as the number of components in the reference GMM. The estimation of the parameters of the target GMM, $p(\mathbf{x}|\boldsymbol{\Theta})$, that best approximates the empirical distribution, $\tilde{p}(\mathbf{x})$, also uses the reference GMM, $p(\mathbf{x}|\tilde{\boldsymbol{\Theta}})$, to form spectral and spatial constraints on the target GMM parameters. Once the target GMM is obtained, the pixels of interest can be selected as the ones that are the most likely under the estimated model.

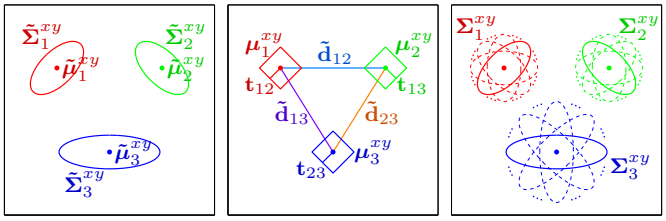
The proposed estimation algorithm is presented in Section IV. The algorithm uses the following constraints that are defined between pairs of Gaussian components, one from the reference GMM and the other one from the target GMM.

- We want to keep the relative sizes of the components of reference and target structures the same, i.e., $\alpha_k = \tilde{\alpha}_k$ for $k = 1, \ldots, K$.
- We want the average spectral content of the reference and target components to be similar. Thus, we constrain the multispectral part of each target mean to lie inside a confidence ellipsoid around the reference mean, i.e., $(\boldsymbol{\mu}_k^{ms} - \tilde{\boldsymbol{\mu}}_k^{ms})^T (\tilde{\boldsymbol{\Sigma}}_k^{ms})^{-1} (\boldsymbol{\mu}_k^{ms} - \tilde{\boldsymbol{\mu}}_k^{ms}) \leq \beta$ for $k = 1, \ldots, K$, where the constant $\beta \in \mathbb{R}_+$ determines the size of the ellipsoid as the tolerance to differences in spectral content. It can also be used to adjust the sensitivity of the model to changes in illumination conditions.
- We also want the homogeneity of the spectral content of the corresponding reference and target components to be the same, i.e., $\boldsymbol{\Sigma}_k^{ms} = \tilde{\boldsymbol{\Sigma}}_k^{ms}$ for $k = 1, \ldots, K$.
- We want to preserve the spatial layout of the reference structure in the target structure. Thus, given the $K(K-$



(a) (b) (c)

Fig. 2. Spectral constraints for an example mixture of three Gaussians. (a) Reference spectral model. (b) Mean constraints: means must lie inside the ellipsoids defined as $(\boldsymbol{\mu}_k^{ms} - \tilde{\boldsymbol{\mu}}_k^{ms})^T(\tilde{\boldsymbol{\Sigma}}_k^{ms})^{-1}(\boldsymbol{\mu}_k^{ms} - \tilde{\boldsymbol{\mu}}_k^{ms}) \leq \beta$. (c) Covariance constraints: $\boldsymbol{\Sigma}_k^{ms} = \tilde{\boldsymbol{\Sigma}}_k^{ms}$.



(a) (b) (c)

Fig. 3. Spatial constraints for an example mixture of three Gaussians. (a) Reference spatial model. (b) Mean constraints: means must lie inside the squares defined as $\boldsymbol{\mu}_i^{xy} + \tilde{\mathbf{d}}_{ij} - \boldsymbol{\mu}_j^{xy} = \mathbf{t}_{ij}$, $\|\mathbf{t}_{ij}\|_1 \leq u$ where $\tilde{\boldsymbol{\mu}}_i^{xy} + \tilde{\mathbf{d}}_{ij} = \tilde{\boldsymbol{\mu}}_j^{xy}$. (c) Covariance constraints: aspect ratios are preserved while rotations are allowed as $\lambda_{min}(\boldsymbol{\Sigma}_k^{xy}) = \lambda_{min}(\tilde{\boldsymbol{\Sigma}}_k^{xy})$ and $\lambda_{max}(\boldsymbol{\Sigma}_k^{xy}) = \lambda_{max}(\tilde{\boldsymbol{\Sigma}}_k^{xy})$.

$1)/2$ displacement vectors $\tilde{\mathbf{d}}_{ij}, i = 1, \ldots, K-1, j = i+1, \ldots, K$, that are computed as in (5), the spatial layout of the target structure is constrained as $\boldsymbol{\mu}_i^{xy} + \tilde{\mathbf{d}}_{ij} - \boldsymbol{\mu}_j^{xy} = \mathbf{t}_{ij}$ where $\|\mathbf{t}_{ij}\|_1 \leq u$ and the constant $u \in \mathbb{R}_+$ specifies the allowed amount of deviation from the reference spatial relations.

- Finally, we want the aspect ratio of each reference primitive object to be preserved in the target. Thus, we constrain the minimum and maximum eigenvalues, λ_{min} and λ_{max}, respectively, of the spatial parts of the reference and target covariances to be the same, i.e., $\lambda_{min}(\boldsymbol{\Sigma}_k^{xy}) = \lambda_{min}(\tilde{\boldsymbol{\Sigma}}_k^{xy})$ and $\lambda_{max}(\boldsymbol{\Sigma}_k^{xy}) = \lambda_{max}(\tilde{\boldsymbol{\Sigma}}_k^{xy})$ for $k = 1, \ldots, K$. Note that this allows different rotations of the primitive objects.

The spectral and spatial constraints are illustrated in Figures 2 and 3, respectively.

IV. Detection Algorithm

The input to the detection problem is the reference GMM, $p(\mathbf{x}|\tilde{\boldsymbol{\Theta}})$, that is estimated from $\tilde{N}$ pixels in the reference compound structure, and a target image containing N pixels, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, among which an unknown subset of size $\tilde{N}$ constitutes the pixels of interest that are assumed to have the empirical distribution, $\tilde{p}(\mathbf{x})$. We do not make any explicit assumption about the distribution of the remaining $N - \tilde{N}$ pixels in the target image.

The goal of the detection algorithm is to estimate the parameters of the target GMM, $p(\mathbf{x}|\boldsymbol{\Theta})$, that best approximates the empirical distribution, $\tilde{p}(\mathbf{x})$. In information theory, the relative entropy or the Kullback-Leibler (KL) divergence [19] is a widely used measure of dissimilarity between two probability distributions $\tilde{p}(\mathbf{x})$ and $p(\mathbf{x})$. It can be interpreted as the additional amount of information required to specify the value of $\mathbf{x}$ as a result of using p instead of the true distribution $\tilde{p}$, and is computed as

$$\begin{aligned} KL(\tilde{p}\|p) &= \int \tilde{p}(\mathbf{x}) \log \frac{\tilde{p}(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} \\ &= \int \tilde{p}(\mathbf{x}) \log \tilde{p}(\mathbf{x}) d\mathbf{x} - \int \tilde{p}(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \\ &= -H(\tilde{p}(\mathbf{x})) - \int \tilde{p}(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{7}$$

where $H(\tilde{p}(\mathbf{x}))$ denotes the differential entropy of the probability distribution $\tilde{p}(\mathbf{x})$. The detection problem in this paper is formulated as the minimization of the KL divergence between $\tilde{p}(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\Theta})$ over the constrained GMM parameters $\boldsymbol{\Theta}$ and the indicator variables $\mathcal{Z} = \{z_1, \ldots, z_N\}$ as

$$\{\boldsymbol{\Theta}^*, \mathcal{Z}^*\} = \arg\min_{\boldsymbol{\Theta}, \mathcal{Z}} KL(\tilde{p}(\mathbf{x})\|p(\mathbf{x}|\boldsymbol{\Theta})). \tag{8}$$

Using (7), we can expand the KL divergence in (8) as

$$\begin{aligned} &KL(\tilde{p}(\mathbf{x})\|p(\mathbf{x}|\boldsymbol{\Theta})) \\ &= -H(\tilde{p}(\mathbf{x})) - \int \frac{1}{\tilde{N}} \sum_{j=1}^{N} z_j \delta(\mathbf{x} - \mathbf{x}_j) \log p(\mathbf{x}|\boldsymbol{\Theta}) d\mathbf{x} \\ &= -H(\tilde{p}(\mathbf{x})) - \frac{1}{\tilde{N}} \sum_{j=1}^{N} z_j \log p(\mathbf{x}_j|\boldsymbol{\Theta}). \end{aligned} \tag{9}$$

Since $H(\tilde{p}(\mathbf{x}))$ does not depend on the parameters $\boldsymbol{\Theta}$ and the indicator variables $\mathcal{Z}$ due to the constraint that they sum to $\tilde{N}$ (the entropy does not change according to which $\tilde{N}$ z_j among N are set to 1), the objective (9) corresponds to the minimization of the negative weighted log-likelihood, or equivalently, maximization of the weighted log-likelihood as

$$\{\boldsymbol{\Theta}^*, \mathcal{Z}^*\} = \arg\max_{\boldsymbol{\Theta}, \mathcal{Z}} \sum_{j=1}^{N} z_j \log p(\mathbf{x}_j|\boldsymbol{\Theta}). \tag{10}$$

Since the objective function in (10) is not jointly concave in $\boldsymbol{\Theta}$ and $\mathcal{Z}$, there is no algorithm that can guarantee to find the global optimum. However, the GMM parameters and the indicator variables that correspond to a local optimum solution of (10) can be obtained via alternating optimization using an expectation-maximization (EM) based algorithm.

The EM algorithm uses distributions over the unobserved label variables to obtain a lower bound for the original log-likelihood function. Let $\mathcal{W} = \{w_{jk} = P(y_j = k|\mathbf{x}_j, \boldsymbol{\Theta}), j = 1, \ldots, N, k = 1, \ldots, K\}$ denote the posterior probabilities of the label variables given the corresponding data points $\mathbf{x}_j, j = 1, \ldots, N$. A lower bound function $F(\mathcal{Z}, \mathcal{W}, \boldsymbol{\Theta})$ for the log-

likelihood function $l(\mathcal{Z}, \mathbf{\Theta})$ can be obtained as

$$\begin{aligned} l(\mathcal{Z}, \mathbf{\Theta}) &= \sum_{j=1}^{N} z_j \log\Big(\sum_{k=1}^{K} \alpha_k p_k(\mathbf{x}_j|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\Big) \\ &= \sum_{j=1}^{N} z_j \log\Big(\sum_{k=1}^{K} w_{jk} \frac{\alpha_k p_k(\mathbf{x}_j|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{w_{jk}}\Big) \\ &\geq \sum_{j=1}^{N} z_j \sum_{k=1}^{K} w_{jk} \log\Big(\frac{\alpha_k p_k(\mathbf{x}_j|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{w_{jk}}\Big) \\ &\qquad \text{(using Jensen's inequality)} \\ &= \sum_{j=1}^{N} z_j \sum_{k=1}^{K} w_{jk}\Big(\log(\alpha_k p_k(\mathbf{x}_j|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)) - \log(w_{jk})\Big) \\ &= F(\mathcal{Z}, \mathcal{W}, \mathbf{\Theta}). \end{aligned} \tag{11}$$

Then, the auxiliary optimization problem that uses the derived lower bound can be written as

$$\begin{aligned} &\text{maximize } F(\mathcal{Z}, \mathcal{W}, \mathbf{\Theta}) \quad \text{over } \mathcal{Z}, \mathcal{W}, \mathbf{\Theta} \\ &\text{subject to } \sum_{k=1}^{K} w_{jk} = 1,\ j = 1, \ldots, N, \\ &\qquad z_j \in \{0, 1\},\ j = 1, \ldots, N, \\ &\qquad \sum_{j=1}^{N} z_j = \tilde{N}, \\ &\qquad \mathbf{\Theta} \in \mathcal{C}_{\mathbf{\Theta}} \end{aligned} \tag{12}$$

where $\mathcal{C}_{\mathbf{\Theta}}$ is the constraint set for $\mathbf{\Theta}$ as defined in Section III. Note that this problem can be solved by introducing a relaxation of the binary indicator variables $z_j \in \{0, 1\}, j = 1, \ldots, N$ as $0 \leq z_j \leq 1$ where an optimal solution still consists of binary values as described below.

The proposed algorithm uses alternating optimization to find a local optimum solution where $F(\mathcal{Z}, \mathcal{W}, \mathbf{\Theta})$ is maximized over $\mathcal{Z}$ for fixed $\mathcal{W}$ and $\mathbf{\Theta}$, and over $\mathcal{W}$ and $\mathbf{\Theta}$ for fixed $\mathcal{Z}$ iteratively. For fixed $\mathcal{W}$ and $\mathbf{\Theta}$, the objective function becomes linear in $\mathcal{Z}$, and maximization of a linear objective over a unit box with a total sum constraint corresponds to a linear program with a simple solution: $z_j = 1$ for the $\tilde{N}$ data points with the largest $\sum_{k=1}^{K} w_{jk}\big(\log(\alpha_k p_k(\mathbf{x}_j|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)) - \log(w_{jk})\big)$ values and $z_j = 0$ for the rest. We call this the Z-step.

For fixed $\mathcal{Z}$, the solutions for $\mathcal{W}$ and $\mathbf{\Theta}$ are similar to the conventional EM algorithm for GMM estimation. If there were no constraints on $\mathbf{\Theta}$, the update equations for the GMM parameters for the optimization of $F(\mathcal{Z}, \mathcal{W}, \mathbf{\Theta})$ via the EM iterations can be derived as

$$w_{jk}^{(t)} = \frac{\alpha_k^{(t)} p_k(\mathbf{x}_j|\boldsymbol{\mu}_k^{(t)}, \mathbf{\Sigma}_k^{(t)})}{\sum_{i=1}^{K} \alpha_i^{(t)} p_i(\mathbf{x}_j|\boldsymbol{\mu}_i^{(t)}, \mathbf{\Sigma}_i^{(t)})} \tag{13}$$

$$\alpha_k^{(t+1)} = \frac{\sum_{j=1}^{N} z_j w_{jk}^{(t)}}{\sum_{i=1}^{K}\sum_{j=1}^{N} z_j w_{ji}^{(t)}} \tag{14}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{j=1}^{N} z_j w_{jk}^{(t)} \mathbf{x}_j}{\sum_{j=1}^{N} z_j w_{jk}^{(t)}} \tag{15}$$

$$\mathbf{\Sigma}_k^{(t+1)} = \frac{\sum_{j=1}^{N} z_j w_{jk}^{(t)} \mathbf{x}_j \mathbf{x}_j^T}{\sum_{j=1}^{N} z_j w_{jk}^{(t)}} - \boldsymbol{\mu}_k^{(t+1)} \big(\boldsymbol{\mu}_k^{(t+1)}\big)^T \tag{16}$$

where (13) corresponds to the E-step, (14)–(16) correspond to the M-step, and the index t corresponds to the iteration number. However, the parameters might not satisfy the desired constraints after being updated in the M-step. Thus, to handle the constraints that are defined with respect to the reference GMM in the previous section, we project the parameters onto constraint sets at the end of every iteration. This can be considered as a special case of the alternating projections algorithm for handling constrained optimization problems [20].

We use the square of the Euclidean distance to measure the distance of a point $\boldsymbol{\theta} \in \Omega$ to a constraint set $\mathcal{C}_{\boldsymbol{\theta}} \subseteq \Omega$ where

$$\text{dist}(\boldsymbol{\theta}, \mathcal{C}_{\boldsymbol{\theta}}) = \min\{\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2^2 \mid \bar{\boldsymbol{\theta}} \in \mathcal{C}_{\boldsymbol{\theta}}\} \tag{17}$$

and Ω is the domain of $\boldsymbol{\theta}$. The point $\bar{\boldsymbol{\theta}} \in \mathcal{C}_{\boldsymbol{\theta}}$ that is closest to $\boldsymbol{\theta}$, i.e., the point for which the minimum in (17) is attained, is referred to as the projection of $\boldsymbol{\theta}$ on $\mathcal{C}_{\boldsymbol{\theta}}$. We use $P_{\mathcal{C}_{\boldsymbol{\theta}}} : \Omega \to \mathcal{C}_{\boldsymbol{\theta}}$ to denote the projection function onto the constraint set $\mathcal{C}_{\boldsymbol{\theta}}$, and $P_{\mathcal{C}_{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ as the projection of $\boldsymbol{\theta}$ on $\mathcal{C}_{\boldsymbol{\theta}}$. Projections $P_{\mathcal{C}_{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ are computed by solving optimization problems defined by the selected constraints. Some of the optimization problems of interest are easy to solve and have simple analytical solutions. However, in general, no analytical solution exists but a solution can be obtained very efficiently using interior point or active set algorithms [21], [22].

The parameters in $\mathbf{\Theta}$ that satisfy the particular constraints defined in Section III can be computed as follows. The prior probabilities can be obtained as the solution of

$$\begin{aligned} &\text{minimize } \sum_{k=1}^{K} |\alpha_k - \tilde{\alpha}_k|^2 \\ &\text{subject to } \alpha_k = \tilde{\alpha}_k,\ k = 1, \ldots, K. \end{aligned} \tag{18}$$

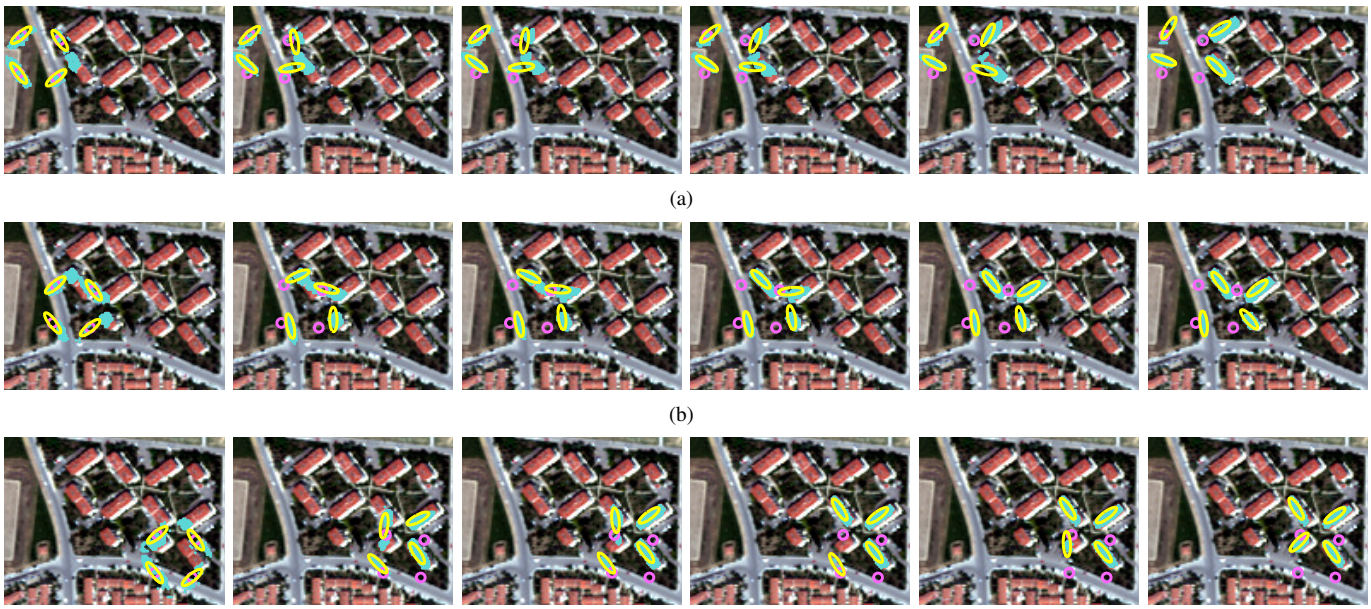
Due to the equality constraints, the optimal value of (18) is achieved with $\alpha_k = \tilde{\alpha}_k$ for $k = 1, \ldots, K$.

Next, the optimization problem and the corresponding projection operator $P_{\mathcal{C}_{\mathbf{\Sigma}}}$ for finding the projections of the covariance matrices is defined as

$$\begin{aligned} &\text{minimize } \sum_{k=1}^{K} \|\mathbf{\Sigma}_k - \tilde{\mathbf{\Sigma}}_k\|_2^2 \\ &\text{subject to } \mathbf{\Sigma}_k^{ms} = \tilde{\mathbf{\Sigma}}_k^{ms},\ k = 1, \ldots, K, \\ &\qquad \lambda_{min}(\tilde{\mathbf{\Sigma}}_k^{xy})\mathbf{I}_2 \leq \mathbf{\Sigma}_k^{xy} \leq \lambda_{max}(\tilde{\mathbf{\Sigma}}_k^{xy})\mathbf{I}_2, \\ &\qquad\qquad k = 1, \ldots, K, \\ &\qquad \mathbf{\Sigma}_k^{i} = 0 \text{ for } i \neq ms, i \neq xy,\ k = 1, \ldots, K \end{aligned} \tag{19}$$

where $\mathbf{I}_2$ is the 2-by-2 identity matrix. The optimal solution for (19) is computed by setting $\mathbf{\Sigma}_k^{ms} = \tilde{\mathbf{\Sigma}}_k^{ms}$, $\mathbf{\Sigma}_k^{i} = 0$ for $i \neq ms$ and $i \neq xy$, and doing eigenvalue decomposition on the spatial part of $\mathbf{\Sigma}_k^{(t+1)}$, thresholding the eigenvalues according to the max-min limits, and reconstructing the spatial part of $\mathbf{\Sigma}_k^{(t+1)}$ using the clipped eigenvalues and eigenvectors.

Finally, the projections of the mean vectors, $P_{\mathcal{C}_{\boldsymbol{\mu}}}$, are computed by solving the following quadratic programming



(a)

(b)

(c)

Fig. 4. Illustration of the convergence of the EM iterations using the reference structure in Figure 5(b). Each row shows a different run with a particular initialization. Each figure shows a particular iteration where the magenta circles are the initial locations of the Gaussians in the first iteration, the cyan dots mark the pixels selected at the end of the Z step, and the yellow ellipses show the current versions of the Gaussians at the end of the M step. The last column shows the solution corresponding to the run in that row. (a) Iterations 1, 16, 32, 48, 64, and 80 are shown. The final log-likelihood was -7103. (b) Iterations 1, 3, 6, 9, 13, and 22 are shown. The final log-likelihood was -5890. (c) Iterations 1, 4, 15, 23, 30, and 38 are shown. The final log-likelihood was -5233. The likelihood values proved to be reliable indicators of the goodness of the solutions with increasing fitness values from (a) to (c).

problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{k=1}^{K} \|\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k\|_2^2 \\ \text{subject to} \quad & (\boldsymbol{\mu}_k^{ms} - \tilde{\boldsymbol{\mu}}_k^{ms})^T (\tilde{\boldsymbol{\Sigma}}_k^{ms})^{-1} (\boldsymbol{\mu}_k^{ms} - \tilde{\boldsymbol{\mu}}_k^{ms}) \leq \beta, \\ & \qquad k = 1, \ldots, K, \\ & \boldsymbol{\mu}_i^{xy} + \tilde{\mathbf{d}}_{ij} - \boldsymbol{\mu}_j^{xy} = \mathbf{t}_{ij}, \ \|\mathbf{t}_{ij}\|_1 \leq u, \\ & \qquad i = 1, \ldots, K-1, j = i+1, \ldots, K. \end{aligned} \tag{20}$$

The proposed detection algorithm is summarized in Algorithm 1. The EM procedure is run by starting from different initializations (described in Section V-B) of the target GMM on the target image. Each run with a particular initialization finds a solution to (12) by alternating between the E, Z, and M steps until an allowed maximum number of iterations is attained or until the difference between the log-likelihood values at two successive iterations falls below some given threshold. The result of each run is the GMM parameters $\boldsymbol{\Theta}^*$ and the indicator variables $\mathcal{Z}^*$ corresponding to a local maximum of the weighted log-likelihood function in (10). Each result involves the binary indicator variables $\{z_j\}_{j=1}^N$ among which $\tilde{N}$ are 1 and $N - \tilde{N}$ are 0, and corresponds to a grouping (selection) of the pixels that have high likelihoods of being similar to the reference Gaussian object model while satisfying the spatial layout constraints. The corresponding likelihood value is considered as a measure of the goodness of that result. The final detection score for each pixel is obtained as the highest likelihood value among the runs in which it is selected. Figure 4 illustrates the convergence of the EM iterations for different initializations.

Algorithm 1 Compound object detection algorithm

Input: $\{\mathbf{x}_j\}_{j=1}^N$, $\{\tilde{\alpha}_k, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k\}_{k=1}^K$, $\{\tilde{\mathbf{d}}_{ij}\}_{i=1,j=i+1}^{K-1,K}$, β, u, $\tilde{N}$
Output: detection scores $\{s_j\}_{j=1}^N$

$s_j \leftarrow -\infty$
for all initializations in the image **do**
  $t \leftarrow 0$
  Set $\alpha_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}$
  repeat {EM iterations}
    E-step: compute $w_{jk}^{(t)}$ using (13)
    Z-step: $z_j \leftarrow 1$ for $\tilde{N}$ data points with largest $\sum_{k=1}^K w_{jk}^{(t)} \big(\log(\alpha_k^{(t)} p_k(\mathbf{x}_j | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})) - \log(w_{jk}^{(t)}) \big)$ and $z_j \leftarrow 0$ for others
    M-step: compute $\alpha_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}$ using (14)–(16)
    Projection: update $\alpha_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}$ by solving (18)–(20)
    $t \leftarrow t + 1$
  until maximum number of iterations reached **or** log-likelihood unchanged
  $s_j \leftarrow \max\{s_j, \textit{log-likelihood}\}, \forall j$ such that $z_j = 1$
end for

V. Experiments

A. Data sets

The experiments were performed using multispectral WorldView-2 images of Ankara and Kusadasi in Turkey. In particular, the Ankara data consisted of a subscene with a size of 700 × 700 pixels and 2 m spatial resolution cover-

ing a residential area with various groups of buildings with different shapes and arrangements as shown in Figure 5(a). The Kusadasi data consisted of two subscenes, each with a size of 600×600 pixels and 2 m spatial resolution, covering also residential areas with different types of building groups as shown in Figures 6(a) and 6(e).

B. Experimental protocol

The input reference compound structures used in the experiments were obtained by manual delineation of the individual primitive objects. This 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. Given a single example structure, the parameters of the reference Gaussian components in $p(\mathbf{x}|\tilde{\mathbf{\Theta}})$ were obtained via maximum likelihood estimation using the pixels belonging to each primitive object. In particular, the component probabilities $\{\tilde{\alpha}_k\}_{k=1}^K$ were estimated using the ratio of the number of pixels in each primitive object to the total number of pixels in the compound structure as in (2), and the means $\{\tilde{\boldsymbol{\mu}}_k\}_{k=1}^K$, the covariance matrices $\{\tilde{\mathbf{\Sigma}}_k\}_{k=1}^K$, and the displacement vectors $\{\tilde{\mathbf{d}}_{ij}\}_{i=1,j=i+1}^{K-1,K}$ were estimated as in (3), (4), and (5), respectively.

After this step with a user input, the rest of the detection process was performed fully unsupervised using the EM algorithm described in Section IV. Note that, the algorithm works on individual pixels without requiring any initial segmentation while performing object detection, but at the same time has the capability of grouping individual pixels that have high likelihoods of belonging to the Gaussian object models while satisfying the spatial layout constraints.

In the EM algorithm, the parameters of the target GMM $p(\mathbf{x}|\mathbf{\Theta})$ were initialized by using the parameters of the reference model. First of all, the number of mixture components (K) was fixed at the number of primitive objects in the reference structure. Next, the Gaussian component probabilities $\{\alpha_k\}_{k=1}^K$ were initialized to the reference Gaussian component probabilities. Similarly, the spectral means $\{\boldsymbol{\mu}_k^{ms}\}_{k=1}^K$ and covariances $\{\mathbf{\Sigma}_k^{ms}\}_{k=1}^K$ were initialized to the reference GMM's corresponding means and covariances.

Since each different initialization of the EM algorithm converges to a local maximum of the likelihood function and there is no prior information about the expected locations of compound structures of interest in the target image, we used a straightforward initialization procedure for the spatial means $\{\boldsymbol{\mu}_k^{xy}\}_{k=1}^K$ using uniform sampling of the image coordinates. A grid of pixels with row and column increments of 20 pixels and a buffer of 30 pixels at the image boundaries was used as offsets to be added to the spatial means of the reference objects while preserving the displacement relations of the spatial means computed from the reference GMM (three examples were given in Figure 4). This resulted in $32 \times 32 = 1024$ runs for the EM algorithm for the Ankara image, and $27 \times 27 = 729$ runs for each of the Kusadasi images. For each run, the spatial covariances $\{\mathbf{\Sigma}_k^{xy}\}_{k=1}^K$ were initialized to the reference GMM's corresponding spatial covariances.

Each EM run solved the optimization problem in (12) with the stopping condition selected to be the difference between the log-likelihood values at two consecutive iterations being less than 10^{-9} or a maximum of 100 iterations. Regarding the parameters in the constraints, β in Figure 2 was set to 10^{-9}, and the deformation parameter u in Figure 3 was set as 0.1 or 10 pixels for a strict or loose satisfaction, respectively, of the spatial layout constraints. Finally, the number of pixels of interest ($\tilde{N}$) was set to the total number of pixels in the reference structure. This choice corresponded to the expectation that the structures of interest in the target image had a similar scale as the reference structure. Detection of structures at scales different from the reference structure is straightforward by scaling $\tilde{N}$ and the parameters of the reference model (spatial means, covariances, and displacement vectors) accordingly. Rotations of the reference structure at 45 and 90 degrees were considered in the experiments below.

C. Baseline for comparison

The baseline method that was used for comparison was the unconstrained Gaussian mixture classifier. The first baseline result (referred to as GMM1) was obtained by computing the likelihood of each pixel using its multispectral values as $\sum_{k=1}^K \tilde{\alpha}_k p_k(\mathbf{x}^{ms}|\tilde{\boldsymbol{\mu}}_k^{ms}, \tilde{\mathbf{\Sigma}}_k^{ms})$ using the GMM estimated from the input reference structure. The second baseline result (referred to as GMM2) was obtained by assigning the highest probability given by individual reference Gaussian components as the detection score of each pixel as $\max_{k=1}^K \tilde{\alpha}_k p_k(\mathbf{x}^{ms}|\tilde{\boldsymbol{\mu}}_k^{ms}, \tilde{\mathbf{\Sigma}}_k^{ms})$. Both of these methods are widely used for GMM-based classification of remotely sensed images in the literature.

D. Evaluation criteria

Thresholding of the detection score at each pixel produces a binary detection map. We used precision and recall as the quantitative performance criteria as in [7] to compare the binary detection maps obtained using a uniformly sampled range of thresholds to the validation data that were obtained by manual labeling of the structures of interest as positive and the rest of the image as negative. Recall (producer's accuracy), that 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), that 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. We also used the F-measure that combines precision and recall using their harmonic mean as

$$\mathrm{F} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{21}$$

to rank different experimental settings and determine the best threshold. A particular threshold value can be selected interactively or by using automatic thresholding techniques [23] when no validation data are available.

In addition to pixel-based evaluation, we also performed object-based evaluation as in [24]. This strategy, 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. In our implementation of the focus-of-attention strategy, we used the convex hull of the pixels belonging to each structure of interest as the object mask for that structure. Then, given the binary detection map for a particular threshold, the union of one or more pixels inside the mask of a target structure was counted as a true positive, while the number of connected components of pixels that did not overlap with any target structure was counted as false positives. Precision and recall used counts of groups of pixels instead of individual pixels for object-based evaluation.

E. Results

The proposed object detection algorithm was evaluated using two scenarios. The first scenario aimed the detection of a structure composed of four buildings with red roofs placed in a diamond formation in the Ankara image as shown in Figure 5. The second scenario aimed the detection of a housing estate composed of four buildings and a pool in the Kusadasi images as shown in Figure 6. Since all three WorldView-2 test images contained suburban scenes, the detection scenarios mainly involved the detection of building groups.

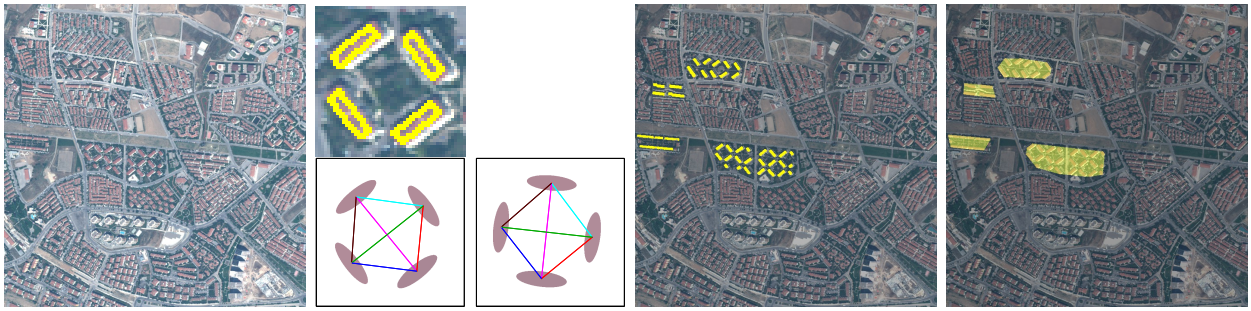
The first set of experiments was done to evaluate the effects of combinations of different rotations of the reference structure in the detection performance as shown in Figure 7. 0 and 45 degree rotations were considered for the Ankara image, whereas 0, 45, and 90 degree rotations were considered for the Kusadasi images as shown in Figures 5 and 6 (0 degree means no rotation). The final detection score for a combination was obtained as the pixelwise maximum of the scores obtained by using individual structures. The results showed that using multiple rotations of the reference structure could improve the performance depending on the image content. For example, the best performance in terms of the F-measure was obtained with the original reference structure (0 degrees) for the Ankara image. However, combining 0 and 90 degree rotations gave the best results for the Kusadasi1 image, and combining 0, 45, and 90 degree rotations gave the best results for the Kusadasi2 image. These results were reasonable considering the appearances of the target structures in the validation data shown in Figures 5 and 6. Note that different rotations of the reference structure affect only the displacement vectors in the layout model as rotations of individual primitive objects were already allowed in the constraints. According to the variance of the F-measure resulting from different combinations, the model with $u = 0.1$ was affected slightly more than the model with $u = 10$ from different combinations. This was also expected because larger amounts of deformations were allowed in the latter model but the former required a more strict satisfaction of the layout constraints.

The next set of experiments was done to compare the performances of the proposed detection algorithm (referred to as CGMM) and the baseline methods (referred to as GMM1 and GMM2) as described in Sections V-B and V-C, respectively. Figure 8 shows the pixel likelihoods as the detection scores for all methods for all images. The best rotation combinations from Figure 7 were used for the CGMM results. The results showed that the proposed algorithm could provide a very good localization of the target structures of interest by incorporating both spectral and structural information in the constrained GMM models. The relative likelihood values were also very strong indicators of the goodness of the detection as the highest likelihood values were obtained for the pixels that belonged to the objects that were very similar to the individual primitives in the reference structures but also satisfied the spatial layout constraints. The spatial constraints that allowed rotations of individual primitives also enabled the detection of additional structures involving cross-like formations or parallel groups of buildings as well as rotated pools while preserving the relative displacements computed from the reference GMMs. On the other hand, the baseline methods that did not use any spatial information detected a wide range of individual objects without any consideration of their spatial arrangements as expected. This led to very low precision and unsatisfactory localization of the structures of interest.

TABLE I
PRECISION, RECALL AND F VALUES FOR THE BEST PERFORMANCE FOR DIFFERENT DETECTION METHODS AND DATA SETS. THE BEST PERFORMANCE CORRESPONDS TO THE LIKELIHOOD THRESHOLD THAT OBTAINED THE HIGHEST F VALUE.

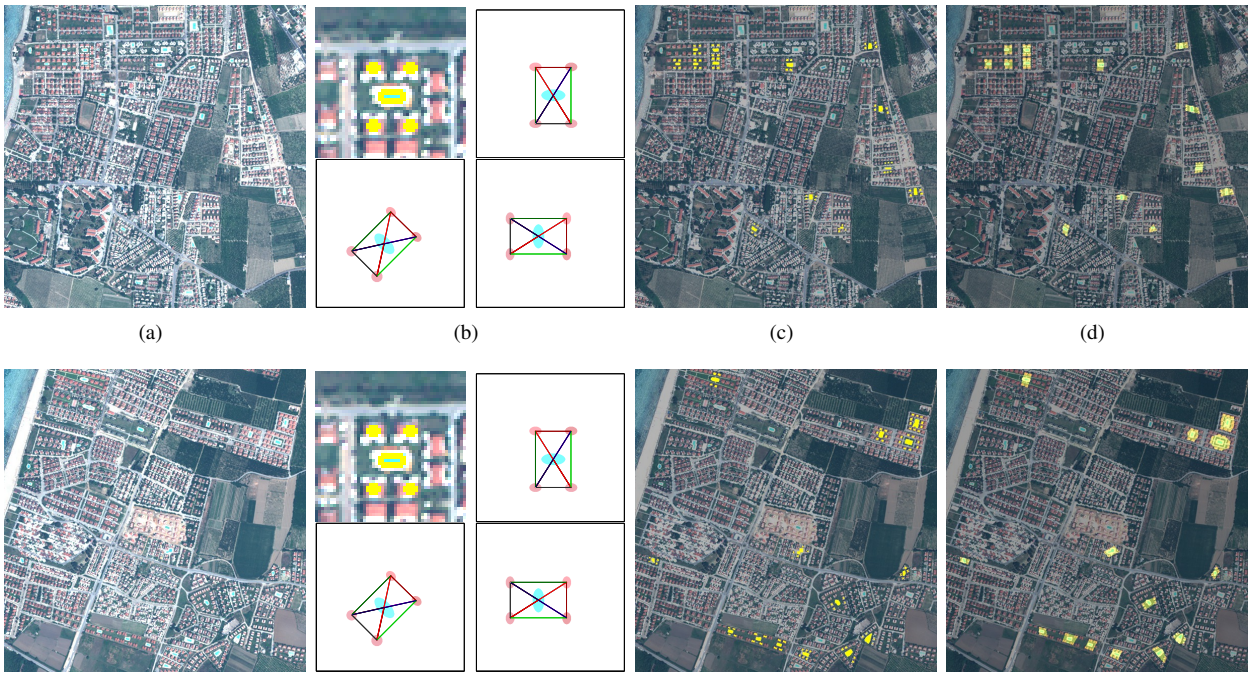
Data	Method	Pixel-based			Object-based		
		Prec.	Rec.	F	Prec.	Rec.	F
Ankara	CGMM ($u=0.1$)	0.9782	0.5292	0.6868	1.0000	0.7500	0.8571
	CGMM ($u=10$)	0.9956	0.6932	0.8173	1.0000	1.0000	1.0000
	GMM1	0.1366	0.3737	0.2000	1.0000	0.2500	0.4000
	GMM2	0.1390	0.4036	0.2068	0.0103	1.0000	0.0204
Kusadasi1	CGMM ($u=0.1$)	0.4619	0.5372	0.4967	0.7692	0.7692	0.7692
	CGMM ($u=10$)	0.4269	0.4783	0.4512	1.0000	0.5385	0.7000
	GMM1	0.0410	0.4175	0.0746	0.0187	0.5385	0.0362
	GMM2	0.0412	0.4104	0.0749	0.0186	0.5385	0.0360
Kusadasi2	CGMM ($u=0.1$)	0.5405	0.5278	0.5341	0.7222	0.9286	0.8125
	CGMM ($u=10$)	0.5071	0.5522	0.5287	0.9231	0.8571	0.8889
	GMM1	0.0949	0.1170	0.1048	0.0267	0.6429	0.0513
	GMM2	0.0846	0.1347	0.1039	0.0269	0.6429	0.0517

Figure 9 shows precision versus recall curves obtained by applying different threshold values to the likelihood based detection scores, and Table I summarizes the results corresponding to the best thresholds. The results showed that the proposed algorithm that jointly exploited spectral and spatial information performed significantly better than the baseline methods that only used spectral information. In particular, CGMM achieved significantly higher precision values than GMM1 and GMM2 for the same value of recall for both pixel-based and object-based evaluation. There was no significant difference between the performances of GMM1 and GMM2, but the accuracies of the models with $u = 0.1$ and $u = 10$ varied according to the appearances of the target structures of interest in different images. For example, the model with $u = 10$ performed significantly better than the one with $u = 0.1$ in the Ankara image because of the flexibility needed in the displacement of the primitive objects in the target structures of interest. However, for the Kusadasi images, the model with $u = 10$ obtained slightly higher precision scores for very high values of recall whereas the model with $u = 0.1$ often had slightly higher precision scores for lower



(a) (b) (c) (d)

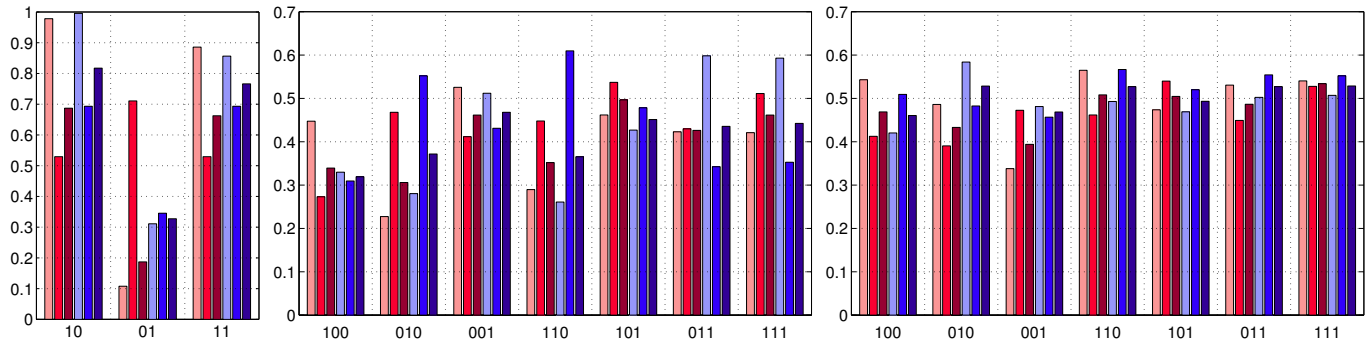
Fig. 5. Target structure composed of four buildings with red roofs in a diamond formation in the Ankara image. (a) RGB image. (b) Close-up (as a 50×50 pixel patch) of the reference structure used in the detection algorithm with the primitive objects overlayed as yellow polygons and the corresponding reference GMMs at 0 and 45 degree rotations. (c) Validation data used for pixel-based evaluation with the individual primitive objects overlayed as yellow. (d) Validation data used for object-based evaluation with the convex hulls of the target structures overlayed as yellow.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6. Target structure composed of a housing estate with four buildings and a pool in the Kusadasi images. (a,e) RGB images (Kusadasi1 and Kusadasi2). (b,f) Close-up (as a 50×50 pixel patch) of the reference structure (from (a)) used in the detection algorithm with the primitive objects overlayed as yellow polygons and the corresponding reference GMMs at 0, 45, and 90 degree rotations. (c,g) Validation data used for pixel-based evaluation with the individual primitive objects overlayed as yellow. (d,h) Validation data used for object-based evaluation with the convex hulls of the target structures overlayed as yellow.



(a) (b) (c)

Fig. 7. Precision, recall, and F values for combinations of different rotations of the reference structure. The binary codes below the plots indicate the rotation settings used for each result: 0 and 45 degrees for the Ankara image in (a), and 0, 45, and 90 degrees for the Kusadasi1 and Kusadasi2 images in (b) and (c), respectively. Precision, recall, and F values (from left to right) are shown as red and blue bars for $u = 0.1$ and $u = 10$, respectively, for each setting.

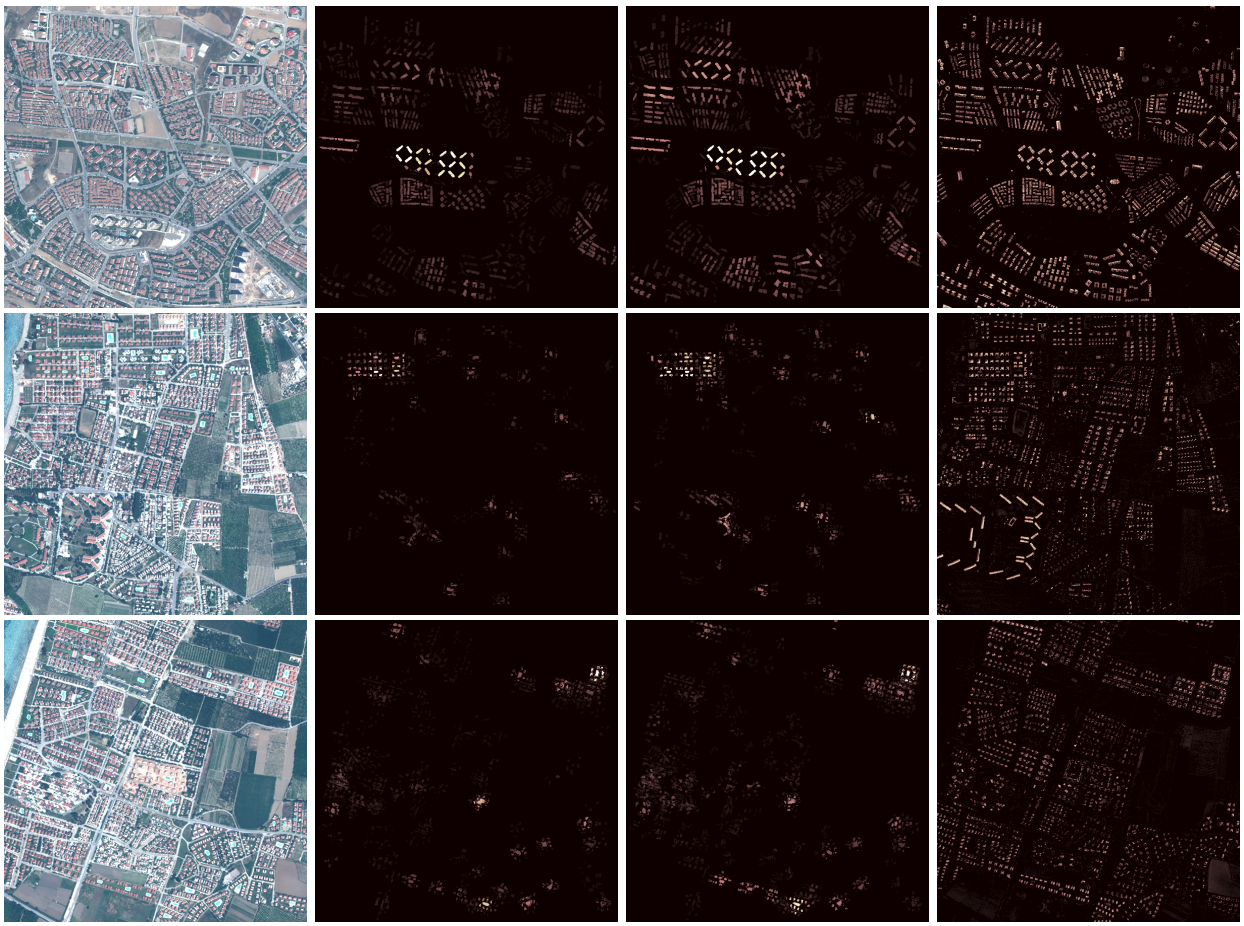


Fig. 8. Pixel likelihoods as detection scores. Brighter values indicate higher likelihoods. The first column shows the RGB images, the second and the third columns show the results for CGMM using $u = 0.1$ and $u = 10$, respectively, and the fourth column shows the results for GMM2. The results for GMM1 were very similar to those of GMM2. Note that the likelihood values for the proposed model were very discriminative and provided good localization.

recall values. This could be explained with the observation that the more strict model ($u = 0.1$) needed to produce more detections to achieve very high recall values but could be more selective when missing some of the targets could be tolerated. The performance scores for the Ankara image were higher in general than the scores for the Kusadasi images because the target primitive objects of interest in the Ankara image had an average size of 120 pixels whereas the target primitives in the Kusadasi images had an average size of 13 pixels, making the latter a much more difficult learning and detection task. The results also showed that object-based performance scores were always higher than those in pixel-based evaluation. This was consistent with the observations in [24] that the object-based target detection evaluation permits a much higher threshold than would be needed to accurately detect most of the pixels in the target, and an increased threshold generally produces fewer false alarms.

Figure 10 shows detection examples. Considering that the primitive objects in the input structures used to estimate the reference GMMs had on average only 80 and 16 pixels for the Ankara (Figure 5) and Kusadasi (Figure 6) images, respectively, the detection results by the proposed method showed a very effective localization of the target structures. For example, even though the spectral-only GMM1 and GMM2 models could not learn and detect the pools in the Kusadasi images, the proposed model could identify most of the pools because of the enhanced likelihood due to the joint use of spectral and spatial information learned from a very small number of pixels. In fact, the proposed model could also allow partial detection of the primitives and showed the ability to handle missing primitives due to the contextual information that it captured even though the decisions were made in the pixel level. Additional constraints can be used to restrict or relax both the appearances and the spatial layout of the primitive objects within the compound structures of interest.

Finally, we analyzed the effect of the deformation parameter on the running time. On the average, one EM run took 17 seconds for $u = 0.1$ and 15.08 seconds for $u = 10$ on a PC with a 2.27 GHz Intel Xeon processor using a Python-based implementation. The running time of our generic implementation of the EM algorithm for constrained GMM estimation using alternating projections could actually be made shorter by exploiting the peculiarities of the specific constraints used. For example, the spectral constraints can be used to reduce the number of EM runs by using a pre-filtering of the image for potential locations of the target structure, and the spatial constraints can be exploited to decrease the search space and reduce the number of pixels used in computing the likelihood.

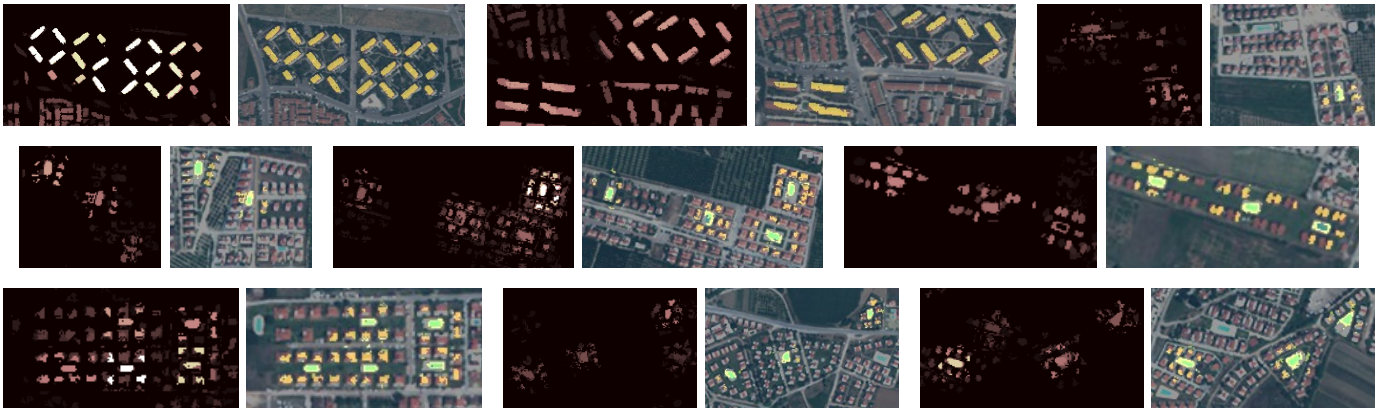


Fig. 10. Examples of local details in the detection results. The image pairs show the likelihood values and the resulting detections after thresholding. The first two rows correspond to CGMM ($u = 10$) and the third row corresponds to CGMM ($u = 0.1$).

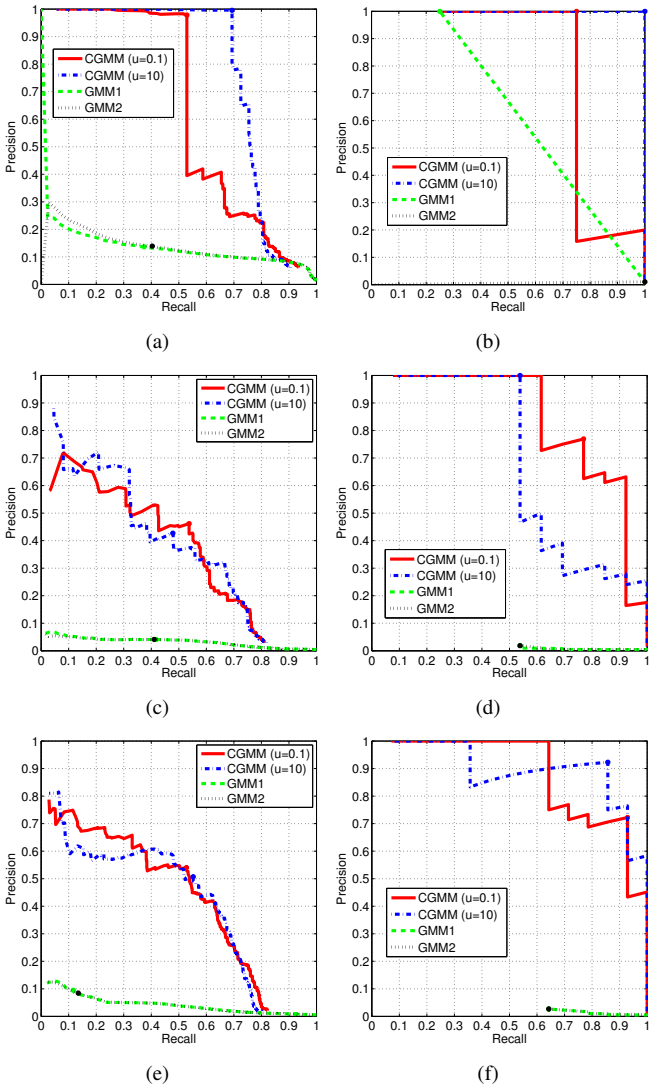


Fig. 9. Precision versus recall curves for CGMM ($u = 0.1$), CGMM ($u = 10$), GMM1, and GMM2 using both pixel-based (a,c,e) and object-based (b,d,f) evaluation. (a,b) Ankara, (c,d) Kusadasi1, (e,f) Kusadasi2 image. The best F value is marked on each curve.

We also observed that the average distance from initialization to convergence was 11 and 18.48 pixels for the Ankara and Kusadasi images, respectively, for $u = 0.1$, and 9.25 and 16.18 pixels for the Ankara and Kusadasi images, respectively, for $u = 10$. This showed that the model with $u = 10$ took both shorter time and shorter distance to converge because of the relaxed constraints. However, it might also require denser initializations to cover a given image space. The analysis showed that the proposed model provided flexibility for possible adjustment of the parameters by the users according to the characteristics of the structures of interest.

VI. Conclusions

We described a new approach for the detection of compound structures that are comprised of spatial arrangements of primitive objects in very high spatial resolution images. The proposed approach used Gaussian mixture models (GMM) to represent the compound structures in which the individual Gaussian components modeled the spectral and shape characteristics of the individual primitives and an associated layout model was used to model their spatial arrangements. Then, a novel expectation-maximization (EM) algorithm that could incorporate spectral and spatial constraints was presented for the estimation of the proposed object representation and the detection of compound structures in new images. Given an example structure, first, a reference GMM and the spatial layout model were estimated from the pixels belonging to the manually delineated primitive objects. Then, the EM algorithm was used to fit a GMM to the target image data so that the pixels that had high likelihoods of belonging to the Gaussian object models and satisfied the spatial layout constraints could be grouped to perform object detection.

The experiments using WorldView-2 images showed that the proposed method could detect high-level structures that cannot be modeled using traditional techniques. The method was capable of very effective localization of the target structures without requiring any image segmentation while performing object detection by grouping individual pixels. Furthermore, the enhanced likelihood computed via the joint use of spectral and spatial information also enabled partial detection of the primitives due to the contextual information that the model

captured from a very small number of example pixels. Future work includes experiments with other types of compound structures in different data sets. We are also planning to extend the model with additional constraints.

REFERENCES

[1] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, March 2013.

[2] S. Aksoy, "Spatial techniques for image classification," in *Signal and Image Processing for Remote Sensing*, C. H. Chen, Ed. CRC Press, 2006, pp. 491–513.

[3] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Multiple spectral-spatial classification approach for hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4122–4132, November 2010.

[4] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial-spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recognition*, vol. 45, no. 1, pp. 381–392, January 2012.

[5] R. R. Vatsavai, B. Bhaduri, A. Cheriyadat, L. Arrowood, E. Bright, S. Gleason, C. Diegert, A. Katsaggelos, T. Pappas, R. Porter, J. Bollinger, B. Chen, and R. Hohimer, "Geospatial image mining for nuclear proliferation detection: Challenges and new opportunities," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, Honolulu, Hawaii, July 25–30, 2010, pp. 48–51.

[6] D. Liu, L. He, and L. Carin, "Airport detection in large aerial optical imagery," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, May 17–21, 2004, pp. 761–764.

[7] S. Aksoy, I. Z. Yalniz, and K. Tasdemir, "Automatic detection and segmentation of orchards using very high-resolution imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, pp. 3117–3131, August 2012.

[8] S. Bhagavathy and B. S. Manjunath, "Modeling and detection of geospatial objects using texture motifs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 12, pp. 3706–3715, December 2006.

[9] N. R. Harvey, C. Ruggiero, N. H. Pawley, B. MacDonald, A. Oyer, L. Balick, and S. P. Brumby, "Detection of facilities in satellite imagery using semi-supervised image classification and auxiliary contextual observables," in *Proceedings of SPIE Visual Information Processing XVIII*, vol. 7341, Orlando, Florida, April 13, 2009.

[10] R. R. Vatsavai, A. Cheriyadat, and S. Gleason, "Supervised semantic classification for nuclear proliferation monitoring," in *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop*, Washington, DC, October 13–15, 2010.

[11] R. Gaetano, G. Scarpa, and G. Poggi, "Hierarchical texture-based segmentation of multiresolution remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2129–2141, July 2009.

[12] D. Zamalieva, S. Aksoy, and J. C. Tilton, "Finding compound structures in images using image segmentation and graph-based knowledge discovery," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, vol. V, Cape Town, South Africa, July 13–17, 2009, pp. 252–255.

[13] H. G. Akcay and S. Aksoy, "Detection of compound structures using multiple hierarchical segmentations," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, July 23–27, 2012, pp. 6833–6836.

[14] K. Grauman and B. Leibe, *Visual Object Recognition*. Morgan & Claypool, 2011.

[15] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, Madison, Wisconsin, June 18–20, 2003, pp. 264–271.

[16] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 259–289, May 2008.

[17] C. Ari and S. Aksoy, "Detection of compound structures using a Gaussian mixture model with spectral and spatial constraints," in *Proceedings of SPIE Defense, Security, and Sensing: Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII*, Baltimore, Maryland, April 23–27, 2012.

[18] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[21] M. S. Andersen, J. Dahl, Z. Liu, and L. Vandenberghe, "Interior-point methods for large-scale cone programming," in *Optimization for Machine Learning*, S. Sra, S. Nowozin, and S. J. Wright, Eds. MIT Press, 2011, pp. 55–83.

[22] M. S. Andersen, J. Dahl, and L. Vandenberghe, "CVXOPT: A Python package for convex optimization," 2012. [Online]. Available: http://abel.ee.ucla.edu/cvxopt

[23] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–165, January 2004.

[24] N. R. Harvey and J. Theiler, "Focus-of-attention strategies for finding discrete objects in multispectral imagery," in *Proceedings of SPIE Imaging Spectrometry X*, vol. 5546, Denver, Colorado, August 2, 2004, pp. 179–189.

Çağlar Arı received his B.S. and Ph.D. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2004 and 2013, respectively. His research interests include machine learning, convex optimization, and computer vision.

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, 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 and the Co-Director of the RETINA Vision and Learning Group. He spent 2013 as a Visiting Associate Professor at the Department of Computer Science & 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 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 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.