# Learning Bayesian Classifiers for Scene Classification With a Visual Grammar

Selim Aksoy, *Member, IEEE*, Krzysztof Koperski, *Member, IEEE*, Carsten Tusk, Giovanni Marchisio, *Member, IEEE*, and James C. Tilton, *Senior Member, IEEE*

*Abstract*—**A challenging problem in image content extraction and classification is building a system that automatically learns high-level semantic interpretations of images. We describe a Bayesian framework for a visual grammar that aims to reduce the gap between low-level features and high-level user semantics. Our approach includes modeling image pixels using automatic fusion of their spectral, textural, and other ancillary attributes; segmentation of image regions using an iterative split-and-merge algorithm; and representing scenes by decomposing them into prototype regions and modeling the interactions between these regions in terms of their spatial relationships. Naive Bayes classifiers are used in the learning of models for region segmentation and classification using positive and negative examples for user-defined semantic land cover labels. The system also automatically learns representative region groups that can distinguish different scenes and builds visual grammar models. Experiments using Landsat scenes show that the visual grammar enables creation of high-level classes that cannot be modeled by individual pixels or regions. Furthermore, learning of the classifiers requires only a few training examples.**

*Index Terms*—**Data fusion, image classification, image segmentation, spatial relationships, visual grammar.**

## I. Introduction

THE AMOUNT of image data that is received from satellites is constantly increasing. For example, the National Aeronautics and Space Administration (NASA) Terra satellite sends more than 850 GB of data to the earth every day (http://terra.nasa.gov). Automatic content extraction, classification, and content-based retrieval have become highly desired goals for developing intelligent databases for effective and efficient processing of remotely sensed imagery. Most of the previous approaches try to solve the content extraction problem by building pixel-based classification and retrieval models using spectral and textural features. However, there is a large semantic gap between low-level features and high-level user expectations and scenarios. This semantic gap makes a human expert's involvement and interpretation in the final analysis inevitable, and this makes processing of data in large remote sensing archives practically impossible.

The commonly used statistical classifiers model image content using distributions of pixels in spectral or other feature domains by assuming that similar land cover structures will cluster together and behave similarly in these feature spaces. Schröder *et al.* [1] developed a system that uses Bayesian classifiers to represent high-level land cover labels for pixels using their low-level spectral and textural attributes. They used these classifiers to retrieve images from remote sensing archives by approximating the probabilities of images belonging to different classes using pixel-level probabilities.

However, an important element of image understanding is the spatial information because complex land cover structures usually contain many pixels and regions that have different feature characteristics. Furthermore, two scenes with similar regions can have very different interpretations if the regions have different spatial arrangements. Even when pixels and regions can be identified correctly, manual interpretation is often necessary for many applications of remote sensing image analysis like land cover classification, urban mapping and monitoring, and ecological analysis in public health studies [2]. These applications will benefit greatly if a system can automatically learn high-level semantic interpretations of scenes instead of classification of only the individual pixels.

The VisiMine system [3] we have developed supports interactive classification and retrieval of remote sensing images by extending content modeling from pixel level to region and scene levels. Pixel-level characterization provides classification details for each pixel with automatic fusion of its spectral, textural, and other ancillary attributes. Following a segmentation process, region-level features describe properties shared by groups of pixels. Scene-level features model the spatial relationships of the regions composing a scene using a visual grammar. This hierarchical scene modeling with a visual grammar aims to bridge the gap between features and semantic interpretation.

This paper describes our work on learning the visual grammar for scene classification. Our approach includes learning prototypes of primitive regions and their spatial relationships for higher level content extraction. Bayesian classifiers that require only a few training examples are used in the learning process. Early work on syntactical description of images includes the picture description language [4] that is based on operators that represent the concatenations between elementary picture components like line segments in line drawings. More advanced image processing and computer vision-based approaches on modeling spatial relationships of regions include using centroid locations and minimum bounding rectangles to compute absolute and relative locations [5]. Centroids and minimum

Manuscript received March 15, 2004; revised September 28, 2004. This work was supported by the National Aeronautics and Space Administration under Contracts NAS5-98053 and NAS5-01123. The VisiMine project was also supported by the U.S. Army under Contracts DACA42-03-C-16 and W9132V-04-C-0001.

S. Aksoy is with the Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey (e-mail: saksoy@cs.bilkent.edu.tr).

K. Koperski, C. Tusk, and G. Marchisio are with Insightful Corporation, Seattle, WA 98109 USA (e-mail: krisk@insightful.com; ctusk@insightful.com; giovanni@insightful.com).

J. C. Tilton is with NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA (e-mail: James.C.Tilton@nasa.gov).

Digital Object Identifier 10.1109/TGRS.2004.839547

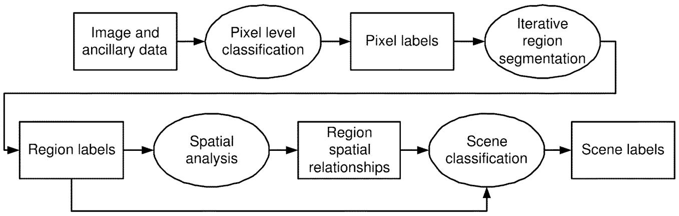


Fig. 1. Object/process diagram for the system. Rectangles represent objects and ellipses represent processes.

bounding rectangles are useful when regions have circular or rectangular shapes, but regions in natural scenes often do not follow these assumptions. More complex representations of spatial relationships include spatial association networks [6], knowledge-based spatial models [7], [8], and attributed relational graphs [9]. However, these approaches require either manual delineation of regions by experts or partitioning of images into grids. Therefore, they are not generally applicable due to the infeasibility of manual annotation in large databases or because of the limited expressiveness of fixed sized grids.

Our work differs from other approaches in that recognition of regions and decomposition of scenes are done automatically after the system learns region and scene models with only a small amount of supervision in terms of positive and negative examples for classes of interest. The rest of the paper is organized as follows. An overview of the visual grammar is given in Section II. The concept of prototype regions is defined in Section III. Spatial relationships of these prototype regions are described in Section IV. Image classification using the visual grammar models is discussed in Section V. Conclusions are given in Section VI.

## II. Visual Grammar

We are developing a visual grammar [10], [11] for interactive classification and retrieval in remote sensing image databases. This visual grammar uses hierarchical modeling of scenes in three levels: pixel level, region level, and scene level. Pixel-level representations include labels for individual pixels computed in terms of spectral features, Gabor [12] and cooccurrence [13] texture features, elevation from digital elevation models (DEMs), and hierarchical segmentation cluster features [14]. Region-level representations include land cover labels for groups of pixels obtained through region segmentation. These labels are learned from statistical summaries of pixel contents of regions using mean, standard deviation, and histograms, and from shape information like area, boundary roughness, orientation, and moments. Scene-level representations include interactions of different regions computed in terms of their spatial relationships.

The object/process diagram of our approach is given in Fig. 1, where rectangles represent objects and ellipses represent processes. The input to the system is raw image and ancillary data. Visual grammar consists of two learning steps. First, pixel-level models are learned using naive Bayes classifiers [1] that provide a probabilistic link between low-level image features and high-level user-defined semantic land cover labels (e.g., city, forest, field). Then, these pixels are combined using an iterative split-and-merge algorithm to find region-level labels. In the second step, a Bayesian framework is used to learn scene classes based on automatic selection of distinguishing spatial relationships between regions. Details of these learning algorithms are given in the following sections. Examples in the rest of the paper use Landsat scenes of Washington, DC, obtained from the NASA Goddard Space Flight Center, and Washington State and Southern British Columbia obtained from the PRISM project at the University of Washington. We use spectral values, Gabor texture features, and hierarchical segmentation cluster features for the first dataset, and spectral values, Gabor features, and DEM data for the second dataset, shown in Fig. 2.

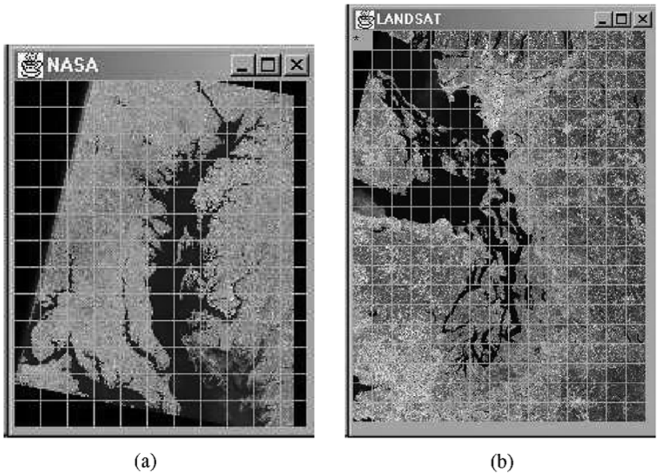


Fig. 2. Landsat scenes used in the experiments. (a) NASA dataset. (b) PRISM dataset.

## III. Prototype Regions

The first step in constructing the visual grammar is to find meaningful and representative regions in an image. Automatic extraction of regions is required to handle large amounts of data. To mimic the identification of regions by analysts, we define the concept of prototype regions. A prototype region is a region that has a relatively uniform low-level pixel feature distribution and describes a simple scene or part of a scene. Ideally, a prototype is frequently found in a specific class of scenes and differentiates this class of scenes from others.

In previous work [10], [11], we used automatic image segmentation and unsupervised model-based clustering to automate the process of finding prototypes. In this paper, we extend this prototype framework to learn prototype models using Bayesian classifiers with automatic fusion of features. Bayesian classifiers allow subjective prototype definitions to be described in terms of easily computable objective attributes. These attributes can be based on spectral values, texture, shape, etc. The Bayesian framework is a probabilistic tool to combine information from multiple sources in terms of conditional and prior probabilities.

Learning of prototypes starts with pixel-level classification (the first process in Fig. 1). Assume there are $k$ prototype labels, $w_1, \ldots, w_k$, defined by the user. Let $x_1, \ldots, x_m$ be the attributes computed for a pixel. The goal is to find the most probable prototype label for that pixel given a particular set of values of these attributes. The degree of association between the pixel and prototype $w_j$ can be computed using the posterior probability

$$\begin{aligned} p(w_j \mid x_1, \ldots, x_m) &= \frac{p(x_1, \ldots, x_m \mid w_j) p(w_j)}{p(x_1, \ldots, x_m)} \\ &= \frac{p(x_1, \ldots, x_m \mid w_j) p(w_j)}{p(x_1, \ldots, x_m \mid w_j) p(w_j) + p(x_1, \ldots, x_m \mid \neg w_j) p(\neg w_j)} \\ &= \frac{p(w_j) \prod_{i=1}^{m} p(x_i \mid w_j)}{p(w_j) \prod_{i=1}^{m} p(x_i \mid w_j) + p(\neg w_j) \prod_{i=1}^{m} p(x_i \mid \neg w_j)} \end{aligned} \quad (1)$$

under the conditional independence assumption. The conditional independence assumption simplifies learning because the parameters for each attribute model $p(x_i \mid w_j)$ can be estimated separately. Therefore, user interaction is only required for the labeling of pixels as positive $(w_j)$ or negative $(\neg w_j)$ examples for a particular prototype label under training. Models for different prototypes are learned separately from the corresponding positive and negative examples. Then, the predicted prototype becomes the one with the largest posterior probability and the pixel is assigned the prototype label

$$w_j^* = \arg \max_{j=1,\ldots,k} p(w_j \mid x_1, \ldots, x_m). \quad (2)$$

We use discrete variables in the Bayesian model where continuous features are converted to discrete attribute values using an unsupervised clustering stage based on the $k$-means algorithm. The number of clusters is empirically chosen for each feature. Clustering is used for processing continuous features (spectral, Gabor, and DEM) and discrete features (hierarchical segmentation clusters) with the same tools. (An alternative is to use a parametric distribution assumption, e.g., Gaussian, for each individual continuous feature, but these parametric assumptions do not always hold.) In the following, we describe learning of the models for $p(x_i \mid w_j)$ using the positive training examples for the $j$th prototype label. Learning of $p(x_i \mid \neg w_j)$ is done the same way using the negative examples.

For a particular prototype, let each discrete variable $x_i$ have $r_i$ possible values (states) with probabilities

$$p(x_i = z \mid \boldsymbol{\theta_i}) = \boldsymbol{\theta_{iz}} > 0 \quad (3)$$

where $z \in \{1, \ldots, r_i\}$, and $\boldsymbol{\theta_i} = \{\boldsymbol{\theta_{iz}}\}_{z=1}^{r_i}$ is the set of parameters for the $i$th attribute model. This corresponds to a multinomial distribution. Since maximum-likelihood estimates can give unreliable results when the sample is small and the number of parameters is large, we use the Bayes estimate of $\boldsymbol{\theta_{iz}}$ that can be computed as the expected value of the posterior distribution.

We can choose any prior for $\boldsymbol{\theta_i}$ in the computation of the posterior distribution, but there is a big advantage to use conjugate priors. A conjugate prior is one which, when multiplied with the direct probability, gives a posterior probability having the same functional form as the prior, thus allowing the posterior to be used as a prior in further computations [15]. The conjugate prior for the multinomial distribution is the Dirichlet distribution [16]. Geiger and Heckerman [17] showed that if all allowed states of the variables are possible (i.e., $\boldsymbol{\theta_{iz}} > 0$) and if certain parameter independence assumptions hold, then a Dirichlet distribution is indeed the only possible choice for the prior.

Given the Dirichlet prior $p(\boldsymbol{\theta_i}) = \mathrm{Dir}(\boldsymbol{\theta_i} \mid \alpha_{i1}, \ldots, \alpha_{ir_i})$ where $\alpha_{iz}$ are positive constants, the posterior distribution of $\boldsymbol{\theta_i}$ can be computed using the Bayes rule as

$$\begin{aligned} p(\boldsymbol{\theta_i} \mid \mathcal{D}) &= \frac{p(\mathcal{D} \mid \boldsymbol{\theta_i}) p(\boldsymbol{\theta_i})}{p(\mathcal{D})} \\ &= \mathrm{Dir}(\boldsymbol{\theta_i} \mid \alpha_{i1} + N_{i1}, \ldots, \alpha_{ir_i} + N_{ir_i}) \end{aligned} \quad (4)$$

where $\mathcal{D}$ is the training sample, and $N_{iz}$ is the number of cases in $\mathcal{D}$ in which $x_i = z$. Then, the Bayes estimate for $\boldsymbol{\theta_{iz}}$ can be found by computing the conditional expected value

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{iz}} = E_{p(\boldsymbol{\theta_i} \mid \mathcal{D})}[\boldsymbol{\theta_{iz}}] = \frac{\alpha_{iz} + N_{iz}}{\alpha_i + N_i} \quad (5)$$

where $\alpha_i = \sum_{z=1}^{r_i} \alpha_{iz}$ and $N_i = \sum_{z=1}^{r_i} N_{iz}$.

An intuitive choice for the hyperparameters $\alpha_{i1}, \ldots, \alpha_{ir_i}$ of the Dirichlet distribution is the Laplace's uniform prior [18] that assumes all $r_i$ states to be equally probable $(\alpha_{iz} = 1, \forall z \in \{1, \ldots, r_i\})$ which results in the Bayes estimate

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{iz}} = \frac{1 + N_{iz}}{r_i + N_i}. \quad (6)$$

Laplace's prior was decided to be a safe choice when the distribution of the source is unknown and the number of possible states $r_i$ is fixed and known [19].

Given the current state of the classifier that was trained using the prior information and the sample $\mathcal{D}$, we can easily update the parameters when new data $\mathcal{D}'$ is available. The new posterior distribution for $\boldsymbol{\theta_i}$ becomes

$$p(\boldsymbol{\theta_i} \mid \mathcal{D}, \mathcal{D}') = \frac{p(\mathcal{D}' \mid \boldsymbol{\theta_i}) p(\boldsymbol{\theta_i} \mid \mathcal{D})}{p(\mathcal{D}' \mid \mathcal{D})}. \quad (7)$$

With the Dirichlet priors and the posterior distribution for $p(\boldsymbol{\theta_i} \mid \mathcal{D})$ given in (4), the updated posterior distribution becomes

$$\begin{aligned} p(\boldsymbol{\theta_i} \mid \mathcal{D}, \mathcal{D}') = \mathrm{Dir}(\boldsymbol{\theta_i} \mid \alpha_{i1} + N_{i1} \\ + N'_{i1}, \ldots, \alpha_{ir_i} + N_{ir_i} + N'_{ir_i}) \end{aligned} \quad (8)$$

where $N'_{iz}$ is the number of cases in $\mathcal{D}'$ in which $x_i = z$. Hence, updating the classifier parameters involves only updating the counts in the estimates for $\hat{\boldsymbol{\theta}}_{\boldsymbol{iz}}$. Figs. 3 and 4 illustrate learning of prototype models from positive and negative examples.

The Bayesian classifiers that are learned as above are used to compute probability maps for all semantic prototype labels and assign each pixel to one of the labels using the maximum a posteriori probability (MAP) rule. In previous work [20], we used a region merging algorithm to convert these pixel-level classification results to contiguous region representations. However, we also observed that this process often resulted in large connected regions and these large regions with very fractal shapes may not be very suitable for spatial relationship computations.

We improved the segmentation algorithm (the second process in Fig. 1) using mathematical morphology operators [21] to automatically divide large regions into more compact subregions. Given the probability maps for all labels where each pixel is assigned either to one of the labels or to the reject class for probabilities smaller than a threshold (latter type of pixels are initially marked as background), the segmentation process proceeds as follows.

1) Merge pixels with identical labels to find the initial set of regions and mark these regions as foreground.

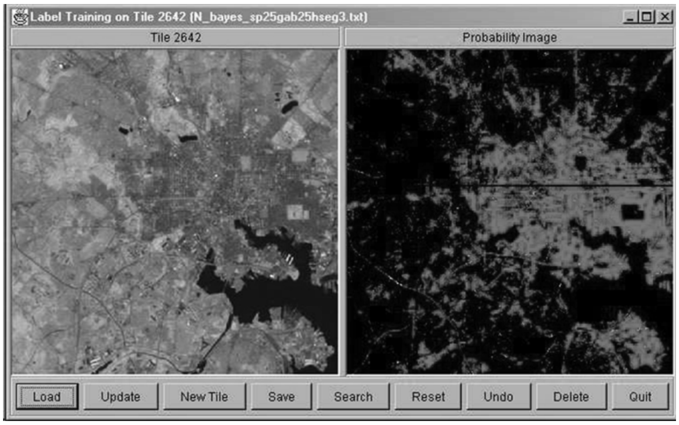


Fig. 3. Training for the city prototype. Positive and negative examples of city pixels in the image on the left are used to learn a Bayesian classifier that creates the probability map shown on the right. Brighter values in the map show pixels with high probability of being part of a city. Pixels marked with red have probabilities above 0.9.

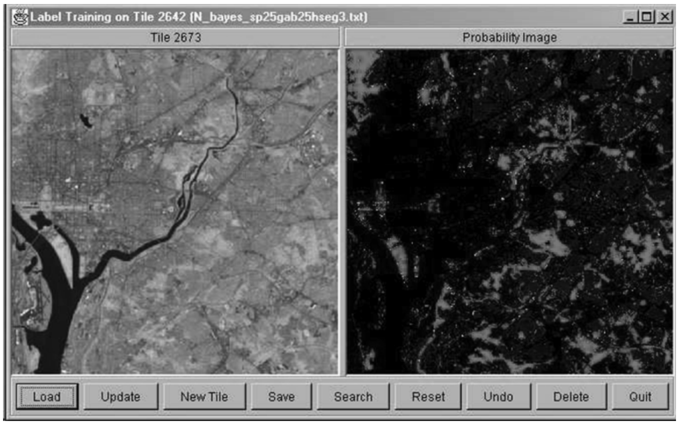
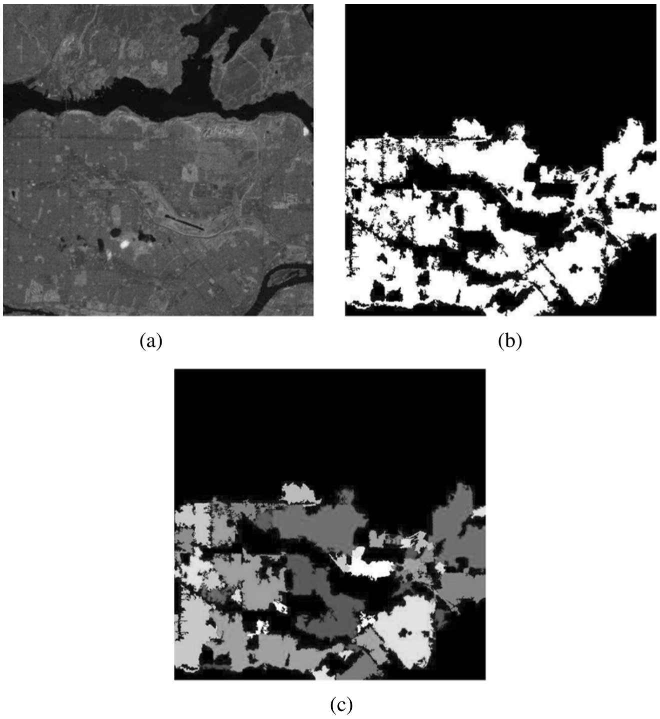


Fig. 4. Training for the park prototype using the process described in Fig. 3.

2) Mark regions with areas smaller than a threshold as background using connected components analysis [21].
3) Use region growing to iteratively assign background pixels to the foreground regions by placing a window at each background pixel and assigning it to the label that occurs the most in its neighborhood.
4) Find individual regions using connected components analysis for each label.
5) For all regions, compute the erosion transform [21] and repeat:
   a) threshold erosion transform at steps of three pixels in every iteration;
   b) find connected components of the thresholded image;
   c) select subregions that have an area smaller than a threshold;
   d) dilate these subregions to restore the effects of erosion;
   e) mark these subregions in the output image by masking the dilation using the original image;

   until no more subregions are found.
6) Merge the residues of previous iterations to their smallest neighbors.


(a) (b)

(c)

Fig. 5. Region segmentation process. The iterative algorithm that uses mathematical morphology operators is used to split a large connected region into more compact subregions. (a) Landsat image, (b) A large connected region formed by merging pixels labeled as residential, (c) More compact subregions.

The merging and splitting process is illustrated in Fig. 5. The probability of each region belonging to a land cover label can be estimated by propagating class labels from pixels to regions. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the set of pixels that are merged to form a region. Let $w_j$ and $p(w_j \mid x_i)$ be the class label and its posterior probability, respectively, assigned to pixel $x_i$ by the classifier. The probability $p(w_j \mid x \in \mathcal{X})$ that a pixel in the merged region belongs to the class $w_j$ can be computed as

$$\begin{aligned} p(w_j \mid x \in \mathcal{X}) &= \frac{p(w_j, x \in \mathcal{X})}{p(x \in \mathcal{X})} \\ &= \frac{p(w_j, x \in \mathcal{X})}{\sum_{t=1}^{k} p(w_t, x \in \mathcal{X})} \\ &= \frac{\sum_{x \in \mathcal{X}} p(w_j, x)}{\sum_{t=1}^{k} \sum_{x \in \mathcal{X}} p(w_t, x)} \\ &= \frac{\sum_{x \in \mathcal{X}} p(w_j \mid x) p(x)}{\sum_{t=1}^{k} \sum_{x \in \mathcal{X}} p(w_t \mid x) p(x)} \\ &= \frac{E_x\{\mathbb{1}_{x \in \mathcal{X}}(x) p(w_j \mid x)\}}{\sum_{t=1}^{k} E_x\{\mathbb{1}_{x \in \mathcal{X}}(x) p(w_t \mid x)\}} \\ &= \frac{1}{n} \sum_{i=1}^{n} p(w_j \mid x_i) \end{aligned} \tag{9}$$

where $\mathbb{1}_A(\cdot)$ is the indicator function associated with the set $A$. Each region in the final segmentation are assigned labels with probabilities using (9).

Fig. 6 shows example segmentations. The number of clusters in $k$-means clustering was empirically chosen as 25 both for spectral values and for Gabor features. The number of clusters for hierarchical segmentation features was automatically obtained as 17. The probability threshold and the minimum

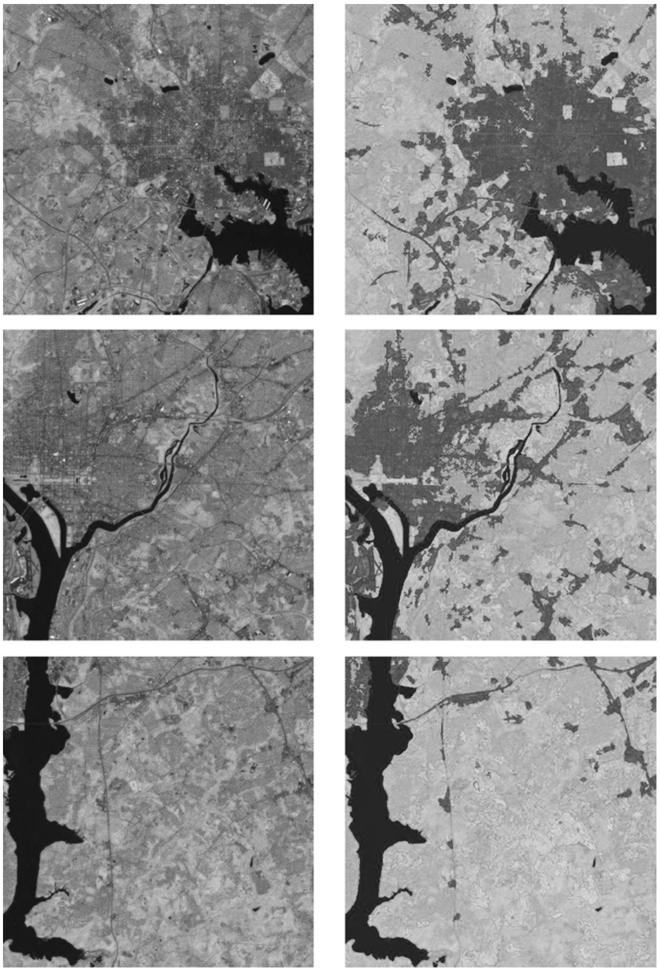


Fig. 6. Segmentation examples from the NASA dataset. Images on the left column are used to train pixel-level classifiers for city, residential area, water, park and field using positive and negative examples for each class. Then, these pixels are combined into regions using the iterative region split-and-merge algorithm and the pixel-level class labels are propagated as labels for these regions. Images on the right column show the resulting region boundaries and the false color representations of their labels for the city (red), residential area (cyan), water (blue), park (green), and field (yellow) classes.

area threshold in the segmentation process were set to 0.2 and 50, respectively. Bayesian classifiers successfully learned proper combinations of features for particular prototypes. For example, using only spectral features confused cities with residential areas and some parks with fields. Using the same training examples, adding Gabor features improved some of the models but still caused some confusion around the borders of two regions with different textures (due to the texture window effects in Gabor computation). We observed that, in general, microtexture analysis algorithms like Gabor features smooth noisy areas and become useful for modeling neighborhoods of pixels by distinguishing areas that may have similar spectral responses but have different spatial structures. Finally, adding hierarchical segmentation features fixed most of the confusions and enabled learning of accurate models from a small set of training examples.

In a large image archive with images of different sensors (optical, hyperspectral, SAR, etc.), training for the prototypes can still be done using the positive and negative examples for each prototype label. If data from more than one sensor are available for the same area, a single Bayes classifier does automatic fusion for a particular label as given in (1) and described above. If different sensors are available for different areas in the same dataset, different classifiers need to be trained for each area (one classifier for each sensor group for each label), again using only positive and negative examples. Once these classifiers that support different sensors for a particular label are trained and the pixels and regions are labeled, the rest of the processes (spatial relationships and image classification) become independent of the sensor data because they use only high-level semantic labels.

## IV. Spatial Relationships

After the images are segmented and prototype labels are assigned to all regions, the next step in the construction of the visual grammar is modeling of region spatial relationships (the third process in Fig. 1). The regions of interest are usually the ones that are close to each other.

Representations of spatial relationships depend on the representations of regions. We model regions by their boundaries. Each region has an outer boundary. Regions with holes also have inner boundaries to represent the holes. Each boundary has a polygon representation of its boundary pixels, and a smoothed polygon approximation, a grid approximation, and a bounding box to speed up polygon intersection operations. In addition, each region has an ID (unique within an image) and a label that is propagated from its pixels' class labels as described in the previous section.

We use fuzzy modeling of pairwise spatial relationships between regions to describe the following high-level user concepts.

*Perimeter-class relationships:*

- *disjoined*: Regions are not bordering each other.
- *bordering*: Regions are bordering each other.
- *invaded_by*: Smaller region is surrounded by the larger one at around 50% of the smaller one's perimeter.
- *surrounded_by*: Smaller region is almost completely surrounded by the larger one.

*Distance-class relationships:*

- *near*: Regions are close to each other.
- *far*: Regions are far from each other.

*Orientation-class relationships:*

- *right*: First region is on the right of the second one.
- *left*: First region is on the left of the second one.
- *above*: First region is above the second one.
- *below*: First region is below the second one.

These relationships are illustrated in Fig. 7. They are divided into subgroups because multiple relationships can be used to describe a region pair at the same time, e.g., *invaded_by* from *left, bordering* from *above*, and *near* and *right*, etc.

To find the relationship between a pair of regions represented by their boundary polygons, we first compute the following:

- perimeter of the first region $\pi_i$;
- perimeter of the second region $\pi_j$;
- common perimeter between two regions $\pi_{ij}$ computed as the shared boundary between two polygons;
- ratio of the common perimeter to the perimeter of the first region $r_{ij} = (\pi_{ij})/(\pi_i)$;

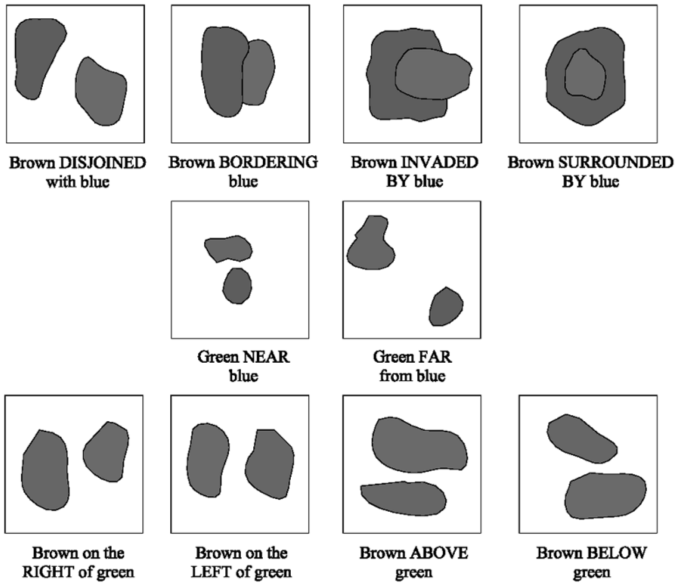
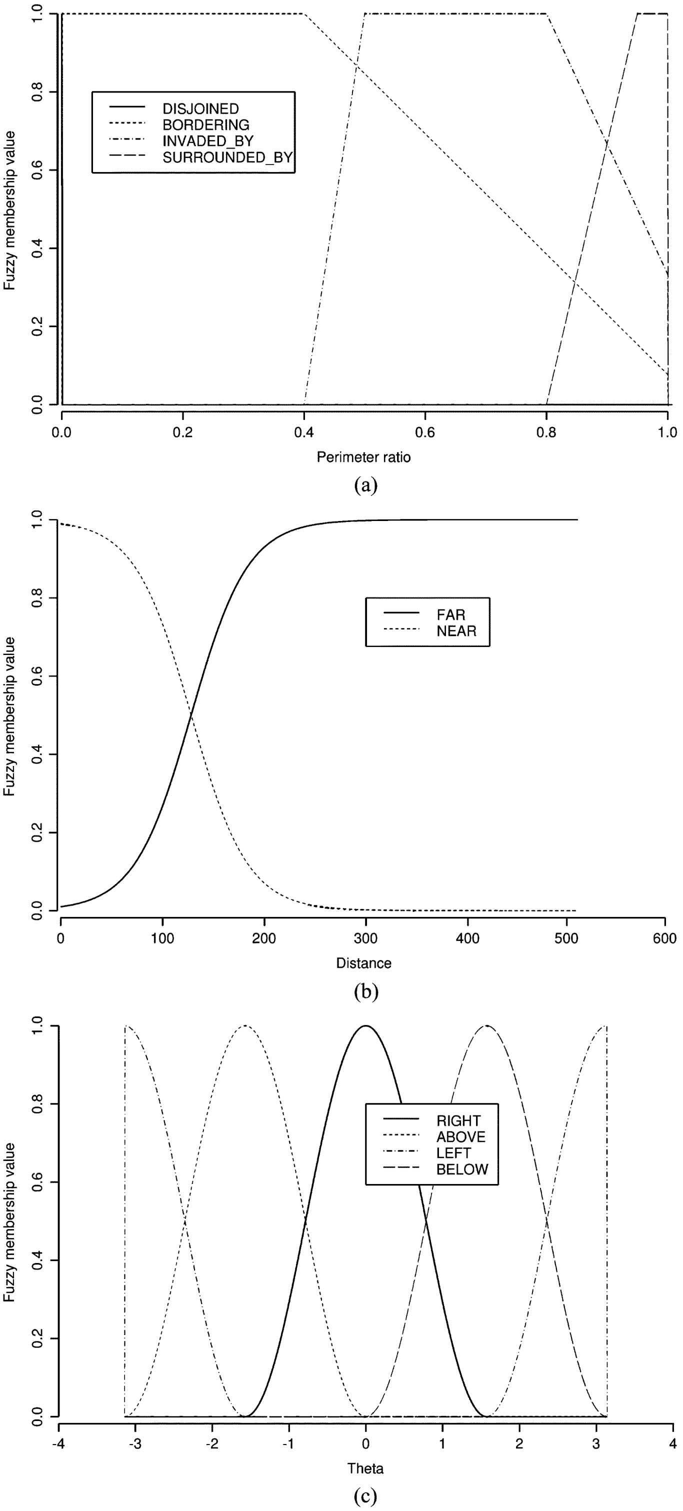


Fig. 7. Spatial relationships of region pairs: *disjoined, bordering, invaded_by, surrounded_by, near, far, right, left, above*, and *below*.



(a)

(b)

(c)

Fig. 8. Fuzzy membership functions for pairwise spatial relationships. (a) Perimeter-class relationships. (b) Distance-class relationships. (c) Orientation-class relationships.

- closest distance between the boundary polygon of the first region and the boundary polygon of the second region $d_{ij}$;
- centroid of the first region $\nu_i$;
- centroid of the second region $\nu_j$;
- angle between the horizontal (column) axis and the line joining the centroids $\theta_{ij}$;

where $i, j \in \{1, \ldots, n\}$ with $n$ being the number of regions in the image. Then, each region pair can be assigned a degree of their spatial relationships using the fuzzy class membership functions given in Fig. 8.

For the perimeter-class relationships, we use the perimeter ratios $r_{ij}$ with trapezoid membership functions. The motivation for the choice of these functions is as follows. Two regions are disjoined when they are not touching each other. They are bordering each other when they have a common boundary. When the common boundary between two regions gets closer to 50%, the larger region starts invading the smaller one. When the common boundary goes above 80%, the relationship is considered an almost complete invasion, i.e., surrounding. For the distance-class relationships, we use the perimeter ratios $r_{ij}$ and boundary polygon distances $d_{ij}$ with sigmoid membership functions. For the orientation-class relationships, we use the angles $\theta_{ij}$ with truncated cosine membership functions. Details of the membership functions are given in [11]. Note that the pairwise relationships are not always symmetric. Furthermore, some relationships are stronger than others. For example, *surrounded_by* is stronger than *invaded_by*, and *invaded_by* is stronger than *bordering*, e.g., the relationship "small region *invaded_by* large region" is preferred over the relationship "large region *bordering* small region." The class membership functions are chosen so that only one of them is the largest for a given set of measurements to avoid ambiguities. The parameters of the functions given in Fig. 8 were manually adjusted to reflect these ideas.

When an area of interest consists of multiple regions, this area is decomposed into multiple region pairs and the measurements defined above are computed for each of the pairwise relationships. Then, these pairwise relationships are combined using an attributed relational graph [21] structure. The attributed relational graph is adapted to our visual grammar by representing regions by the graph nodes and their spatial relationships by the edges between such nodes. Nodes are labeled with the class (land cover) names and the corresponding confidence values (posterior probabilities) for these class assignments. Edges are

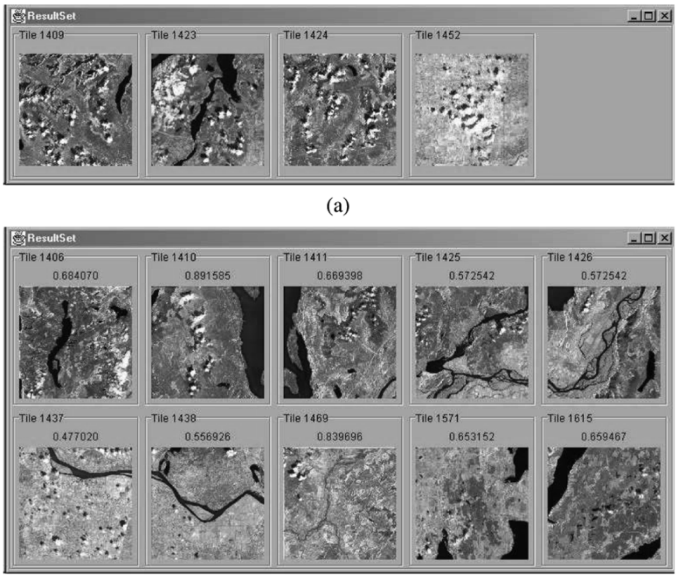


Fig. 9. Classification results for the "clouds" class which is automatically modeled by the distinguishing relationships of white regions (clouds) with their neighboring dark regions (shadows). (a) Training images for clouds. (b) Images classified as containing clouds.

labeled with the spatial relationship classes (pairwise relationship names) and the corresponding degrees (fuzzy membership values) for these relationships.

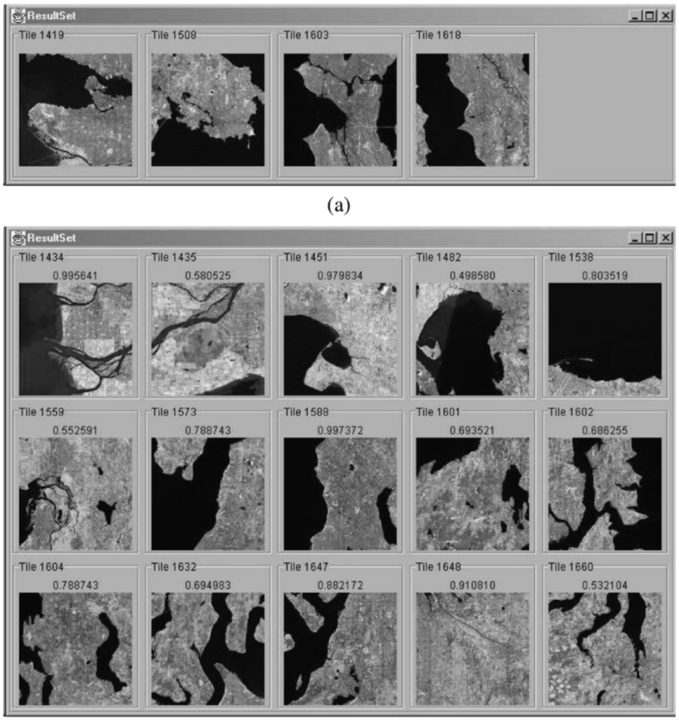


Fig. 10. Classification results for the "residential areas with a coastline" class which is automatically modeled by the distinguishing relationships of regions containing a mixture of concrete, grass, trees, and soil (residential areas) with their neighboring blue regions (water). (a) Training images for residential areas with a coastline. (b) Images classified as containing residential areas with a coastline.

## V. Image Classification

Image classification is defined here as a problem of assigning images to different classes according to the scenes they contain (the last process in Fig. 1). The visual grammar enables creation of high-level classes that cannot be modeled by individual pixels or regions. Furthermore, learning of these classes require only a few training images. We use a Bayesian framework that learns scene classes based on automatic selection of distinguishing (e.g., frequently occurring, rarely occurring) region groups.

The input to the system is a set of training images that contain example scenes for each class defined by the user. Denote these classes by $w_1, \ldots, w_s$. Our goal is to find representative region groups that describe these scenes. The system automatically learns classifiers from the training data as follows.

1) Count the number of times each possible region group (combinatorially formed using all possible relationships between all possible prototype regions) is found in the set of training images for each class. A region group of interest is the one that is frequently found in a particular class of scenes but rarely exists in other classes. For each region group, this can be measured using class separability which can be computed in terms of within-class and between-class variances of the counts as

$$\varsigma = \log\left(1 + \frac{\sigma_B^2}{\sigma_W^2}\right) \tag{10}$$

where $\sigma_W^2 = \sum_{i=1}^{s} v_i \mathrm{var}\{z_j \mid j \in w_i\}$ is the within-class variance, $v_i$ is the number of training images for class $w_i$, $z_j$ is the number of times this region group is found in training image $j$, $\sigma_B^2 = \mathrm{var}\{\sum_{j \in w_i} z_j \mid i = 1, \ldots, s\}$ is the between-class variance, and $\mathrm{var}\{\cdot\}$ denotes the variance of a sample.

2) Select the top $t$ region groups with the largest class separability values. Let $x_1, \ldots, x_t$ be Bernoulli random variables[1] for these region groups, where $x_j = T$ if the region group $x_j$ is found in an image and $x_j = F$ otherwise. Let $p(x_j = T) = \theta_j$. Then, the number of times $x_j$ is found in images from class $w_i$ has a $\mathrm{Binomial}(v_i, \theta_j) = (\begin{smallmatrix} v_i \\ v_{ij} \end{smallmatrix})\theta_j^{v_{ij}}(1-\theta_j)^{v_i - v_{ij}}$ distribution where $v_{ij}$ is the number of training images for $w_i$ that contain $x_j$. Using a Beta(1, 1) distribution as the conjugate prior, the Bayes estimate for $\theta_j$ becomes

$$p(x_j = T \mid w_i) = \frac{v_{ij} + 1}{v_i + 2}. \tag{11}$$

Using a similar procedure with multinomial distributions and Dirichlet priors, the Bayes estimate for an image belonging to class $w_i$ (i.e., containing the scene defined by class $w_i$) is computed as

$$p(w_i) = \frac{v_i + 1}{\sum_{i=1}^{s} v_i + s}. \tag{12}$$

3) For an unknown image, search for each of the $t$ region groups (determine whether $x_j = T$ or $x_j = F$, $\forall j$) and assign that image to the best matching class using the

[1] Finding a region group in an image can be modeled as a Bernoulli trial because there are only two outcomes: the region group is either in the image or not.

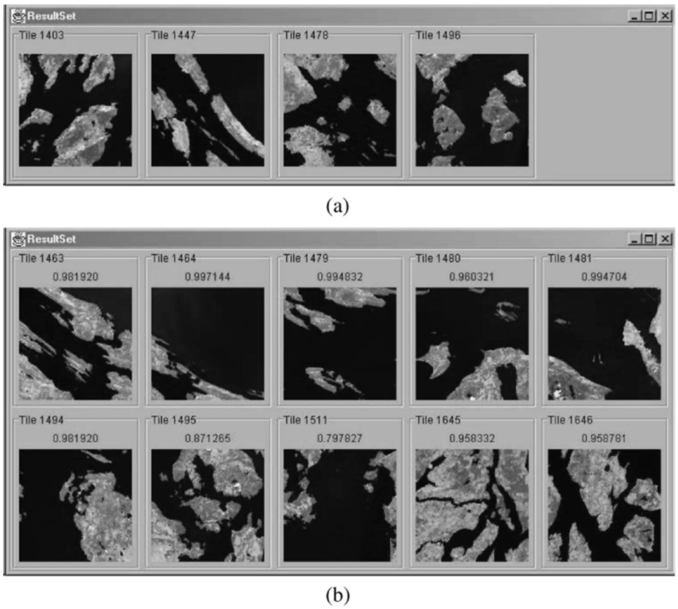


Fig. 11. Classification results for the "tree-covered islands" class which is automatically modeled by the distinguishing relationships of green regions (lands covered with conifer and deciduous trees) surrounded by blue regions (water). (a) Training images for tree-covered islands. (b) Images classified as containing tree-covered islands.

MAP rule with the conditional independence assumption as

$$\begin{aligned} w^* &= \arg\max_{w_i} p(w_i \mid x_1, \ldots, x_t) \\ &= \arg\max_{w_i} p(w_i) \prod_{j=1}^{t} p(x_j \mid w_i). \end{aligned} \tag{13}$$

Classification examples from the PRISM dataset that includes 299 images are given in Figs. 9–11. In these examples, we used four training images for each of the six classes defined as "clouds," "residential areas with a coastline," "tree-covered islands," "snow-covered mountains," "fields," and "high-altitude forests." Commonly used statistical classifiers require a lot of training data to effectively compute the spectral and textural signatures for pixels and also cannot do classification based on high-level user concepts because of the lack of spatial information. Rule-based classifiers also require significant amount of user involvement every time a new class is introduced to the system. The classes listed above provide a challenge where a mixture of spectral, textural, elevation and spatial information is required for correct identification of the scenes. For example, pixel-level classifiers often misclassify clouds as snow and shadows as water. On the other hand, the Bayesian classifier described above can successfully eliminate most of the false alarms by first recognizing regions that belong to cloud and shadow prototypes and then verify these region groups according to the fact that clouds are often accompanied by their shadows in a Landsat scene. Other scene classes like residential areas with a coastline or tree-covered islands cannot be identified by pixel-level or scene-level algorithms that do not use spatial information. While quantitative comparison of results would be difficult due to the unavailability of ground truth for high-level semantic classes for this archive, our qualitative evaluation showed that the visual grammar classifiers automatically learned the distinguishing region groups that were frequently found in particular classes of scenes but rarely existed in other classes.

## VI. Conclusion

We described a visual grammar that aims to bridge the gap between low-level features and high-level semantic interpretation of images. The system uses naive Bayes classifiers to learn models for region segmentation and classification from automatic fusion of features, fuzzy modeling of region spatial relationships to describe high-level user concepts, and Bayesian classifiers to learn image classes based on automatic selection of distinguishing (e.g., frequently occurring, rarely occurring) relations between regions.

The visual grammar overcomes the limitations of traditional region- or scene-level image analysis algorithms which assume that the regions or scenes consist of uniform pixel feature distributions. Furthermore, it can distinguish different interpretations of two scenes with similar regions when the regions have different spatial arrangements. The system requires only a small amount of training data expressed as positive and negative examples for the classes defined by the user. We demonstrated our system with classification scenarios that could not be handled by traditional pixel-, region-, or scene-level approaches but where the visual grammar provided accurate and effective models.

## References

[1] M. Schroder, H. Rehrauer, K. Siedel, and M. Datcu, "Interactive learning and probabilistic retrieval in remote sensing image archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 5, pp. 2288–2298, Sep. 2000.

[2] S. I. Hay, M. F. Myers, N. Maynard, and D. J. Rogers, Eds., "Special Issue: From Remote Sensing to Relevant Sensing in Human Health," *Photogramm. Eng. Remote Sens.*, vol. 68, no. 2, Feb. 2002.

[3] K. Koperski, G. Marchisio, S. Aksoy, and C. Tusk, "VisiMine: Interactive mining in image databases," in *Proc. IGARSS*, vol. 3, Toronto, ON, Canada, Jun. 2002, pp. 1810–1812.

[4] A. C. Shaw, "Parsing of graph-representable pictures," *J. ACM*, vol. 17, no. 3, pp. 453–481, Jul. 1970.

[5] J. R. Smith and S.-F. Chang, "VisualSEEk: A fully automated content-based image query system," in *Proc. ACM Int. Conf. Multimedia*, Boston, MA, Nov. 1996, pp. 87–98.

[6] P. J. Neal, L. G. Shapiro, and C. Rosse, "The digital anatomist structural abstraction: A scheme for the spatial description of anatomical entities," in *Proc. Amer. Medical Informatics Assoc. Annu. Symp.*, Lake Buena Vista, FL, Nov. 1998.

[7] W. W. Chu, C.-C. Hsu, A. F. Cardenas, and R. K. Taira, "Knowledge-based image retrieval with spatial and temporal constructs," *IEEE Trans. Knowl. Data Eng.*, vol. 10, no. 6, pp. 872–888, Nov./Dec. 1998.

[8] L. H. Tang, R. Hanka, H. H. S. Ip, and R. Lam, "Extraction of semantic features of histological images for content-based retrieval of images," in *Proc. SPIE Conf. Medical Imaging*, vol. 3662, San Diego, CA, Feb. 1999, pp. 360–368.

[9] E. G. M. Petrakis and C. Faloutsos, "Similarity searching in medical image databases," *IEEE Trans. Knowl. Data Eng.*, vol. 9, no. 3, pp. 435–447, May/Jun. 1997.

[10] S. Aksoy, G. Marchisio, K. Koperski, and C. Tusk, "Probabilistic retrieval with a visual grammar," in *Proc. IGARSS*, vol. 2, Toronto, ON, Canada, Jun. 2002, pp. 1041–1043.

[11] S. Aksoy, C. Tusk, K. Koperski, and G. Marchisio, "Scene modeling and image mining with a visual grammar," in *Frontiers of Remote Sensing Information Processing*, C. H. Chen, Ed. Singapore: World Scientific, 2003, pp. 35–62.

[12] G. M. Haley and B. S. Manjunath, "Rotation-invariant texture classification using a complete space-frequency model," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 255–269, Feb. 1999.

[13] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[14] J. C. Tilton, G. Marchisio, K. Koperski, and M. Datcu, "Image information mining utilizing hierarchical segmentation," in *Proc. IGARSS*, vol. 2, Toronto, ON, Canada, Jun. 2002, pp. 1029–1031.

[15] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[16] M. H. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[17] D. Geiger and D. Heckerman, "A characterization of the Dirichlet distribution through global and local parameter independence," *Ann. Stat.*, vol. 25, no. 3, pp. 1344–1369, 1997.

[18] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[19] R. F. Krichevskiy, "Laplace's law of succession and universal encoding," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 296–303, Jan. 1998.

[20] S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, and J. C. Tilton, "Learning Bayesian classifiers for a visual grammar," in *Proc. IEEE GRSS Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, Washington, DC, Oct. 2003, pp. 212–218.

[21] R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*. Reading, MA: Addison-Wesley, 1992.

**Selim Aksoy** (S'96–M'01) received the B.S. degree from 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, all in electrical engineering.

He is currently an Assistant Professor in the Department of Computer Engineering, Bilkent University, Ankara. Before joining Bilkent, he was a Research Scientist with 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. From 1996 to 2001, he was a Research Assistant with the University of Washington where he developed algorithms for content-based image retrieval, statistical pattern recognition, object recognition, graph-theoretic clustering, relevance feedback, and mathematical morphology. During summers of 1998 and 1999, he was a Visiting Researcher with the Tampere International Center for Signal Processing collaborating in a content-based multimedia retrieval project. His research interests are in 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 International Association for Pattern Recognition (IAPR). He was recently elected the Vice Chair of the IAPR Technical Committee on Remote Sensing for the period 2004–2006.

**Krzysztof Koperski** (S'88–M'90) received the M.Sc. degree in electrical engineering from Warsaw University of Technology, Warsaw, Poland, in 1989, and the Ph.D. degree in computer science from Simon Fraser University, Burnaby, BC, Canada, in 1999. During his graduate work at Simon Fraser University, he worked on knowledge discovery in spatial databases and spatial data warehousing.

In 1999, he was a Visiting Researcher with the University of L'Aquila, working on spatial data mining in the presence of uncertain information. Since 1999, he has been with Insightful Corporation, Seattle, WA. His research interests include spatial and image data mining, information visualization, information retrieval, and text mining. He has been involved in projects concerning remote sensing image classification, medical image processing, data clustering, and natural language processing.

**Carsten Tusk** received the diploma degree in computer science and engineering from the Rhineland-Westphalian Technical University, Aachen, Germany, in 2001.

Since August 2001, he has been working as a Research Scientist with Insightful Corporation, Seattle, WA. His research interests are in information retrieval, statistical data analysis, and database systems.

**Giovanni Marchisio** received the B.A.Sc. degree in engineering from the University of British Columbia, Vancouver, BC, Canada, and the Ph.D. degree in geophysics and planetary physics from the Scripps Institution of Oceanography, University of California, San Diego.

He is currently Director of Emerging Products with Insightful Corporation, Seattle, WA. He has more than 15 years experience in commercial software development related to text analysis, computational linguistics, image processing, and multimedia information retrieval methodologies. At Insightful, he has been a Principal Investigator on R&D government contracts totaling several millions (with the National Aeronautics and Space Administration, the National Institutes of Health, the Defense Advanced Research Projects Agency, and the Department of Defense). He has articulated novel scientific ideas and software architectures in the areas of artificial intelligence, pattern recognition, Bayesian and multivariate inference, latent semantic analysis, cross-language retrieval, satellite-image-mining World Wide Web-based environment for video and image compression and retrieval. He has also been a Senior Consultant on statistical modeling and prediction analysis of very large databases of multivariate time series. He has authored and coauthored several articles or book chapters on multimedia data mining. In the past three years, he has produced several inventions for information retrieval and knowledge discovery, which led to three U.S. patents. He also has also been a Visiting Professor with the University of British Columbia. His previous work includes research in signal processing, ultrasound acoustic imaging, seismic reflection, and electromagnetic induction imaging of the earth's interior.

**James C. Tilton** (S'79–M'81–SM'94) received the B.A. degrees in electronic engineering, environmental science and engineering, and anthropology, the M.E.E. degree in electrical engineering from Rice University, Houston, TX, in 1976, the M.S. degree in optical sciences from the University of Arizona, Tucson, in 1978, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1981.

He is currently a Computer Engineer with the Applied Information Science Branch (AISB), Earth and Space Data Computing Division, NASA Goddard Space Flight Center, Greenbelt, MD. He was previously with the Computer Sciences Corporation, Silver Spring, MD, from 1982 to 1983, and Science Applications Research, Riverdale, MD, from 1983 to 1985 on contracts with NASA Goddard. As a member of the AISB, he is responsible for designing and developing computer software tools for space and earth science image analysis algorithms, and encouraging the use of these computer tools through interactions with space and earth scientists. His development of a recursive hierarchical segmentation algorithm has resulted in two patent applications. He is an Associate Editor for *Pattern Recognition*.

Dr. Tilton is a member of the IEEE Geoscience and Remote Sensing and Signal Processing Societies and is a member of Phi Beta Kappa, Tau Beta Pi, and Sigma Xi. From 1992 through 1996, he served as a member of the IEEE Geoscience and Remote Sensing Society Administrative Committee. Since 1996, he has served as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.