Learning Bayesian Classifiers for Scene Classification with a Visual Grammar

Selim Aksoy, *Member, IEEE,* Krzysztof Koperski, *Member, IEEE,* Carsten Tusk, Giovanni Marchisio, *Member, IEEE,* 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*— Image classification, visual grammar, image segmentation, spatial relationships, data fusion**

I. Introduction

THE amount of image data that is received from satellites is constantly increasing. For example, NASA's Terra satellite sends more than 850GB of data to Earth every day [1]. 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.* [2] 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 [3]. 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 [4] 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 [5] 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 [6]. Centroids and minimum bounding rectangles are useful

Manuscript received March 15, 2004; revised September 9, 2004. This work was supported by NASA contracts NAS5-98053 and NAS5-01123. VisiMine project was also supported by the U.S. Army contracts DACA42-03-C-16 and W9132V-04-C-0001.

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

K. Koperski, C. Tusk and G. Marchisio are with Insightful Corporation, 1700 Westlake Ave. N., Suite 500, Seattle, WA, 98109, USA. Email: {krisk,ctusk,giovanni}@insightful.com.

J. C. Tilton is with NASA Goddard Space Flight Center, Mail Code 935, Greenbelt, MD, 20771, USA. Email: James.C.Tilton@nasa.gov.

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 [7], knowledge-based spatial models [8], [9], and attributed relational graphs [10]. 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 [11], [12] 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 [13] and co-occurrence [14] texture features, elevation from DEM, and hierarchical segmentation cluster features [15]. 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 [2] 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, D.C., 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 data set, and spectral values, Gabor features and DEM data for the second data set, shown in Fig. 2.

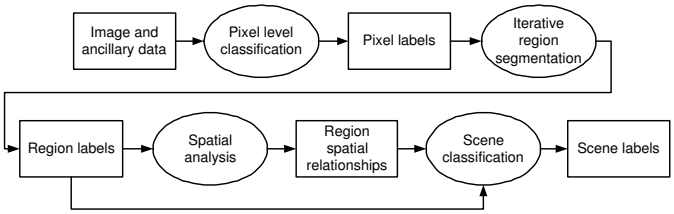
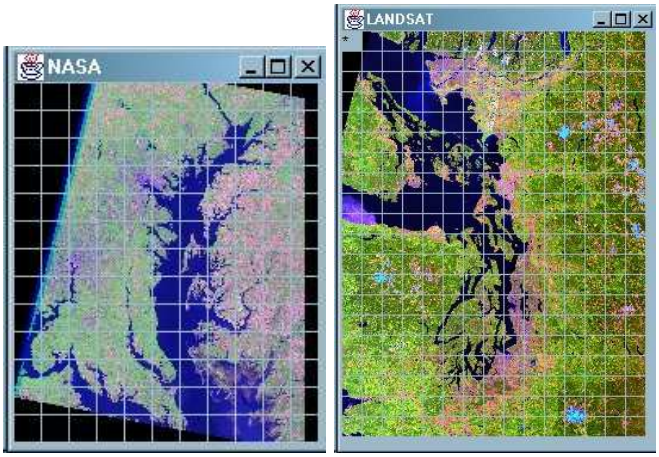


Fig. 1. Object/process diagram for the system. Rectangles represent objects and ellipses represent processes.



(a) NASA data set (b) PRISM data set

Fig. 2. LANDSAT scenes used in the experiments.

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 [11], [12], 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. 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

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

under the conditional independence assumption. The conditional independence assumption simplifies learning because the parameters for each attribute model $p(x_i|w_j)$ 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

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

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 $p(x_i|w_j)$ using the positive training examples for the j'th prototype label. Learning of $p(x_i|\neg w_j)$ 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

$$p(x_i = z|\boldsymbol{\theta_i}) = \boldsymbol{\theta}_{\boldsymbol{i}z} > 0 \tag{3}$$

where $z \in \{1,\ldots,r_i\}$ and $\boldsymbol{\theta_i} = \{\boldsymbol{\theta}_{\boldsymbol{i}z}\}_{z=1}^{r_i}$ 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 $\boldsymbol{\theta}_{\boldsymbol{i}z}$ 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 [16]. The conjugate prior for the multinomial distribution is the Dirichlet distribution [17]. Geiger and Heckerman [18] showed that if all allowed states of the variables are possible (i.e., $\boldsymbol{\theta}_{\boldsymbol{i}z} > 0$) and if certain parameter independence assumptions hold, then a Dirichlet distribution is indeed the only possible choice for the prior.

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

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

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 $\boldsymbol{\theta}_{\boldsymbol{i}z}$ can be found by taking the conditional expected value

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{i}z} = E_{p(\boldsymbol{\theta_i}|\mathcal{D})}[\boldsymbol{\theta}_{\boldsymbol{i}z}] = \frac{\alpha_{iz}+N_{iz}}{\alpha_i+N_i} \tag{5}$$

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 hyper-parameters $\alpha_{i1},\ldots,\alpha_{ir_i}$ of the Dirichlet distribution is the Laplace's uniform prior [19] 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

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{i}z} = \frac{1+N_{iz}}{r_i+N_i}. \tag{6}$$

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 [20].

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

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

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

$$p(\boldsymbol{\theta_i}|\mathcal{D},\mathcal{D}') = \text{Dir}(\boldsymbol{\theta_i}|\alpha_{i1}+N_{i1}+N'_{i1},\ldots,\alpha_{ir_i}+N_{ir_i}+N'_{ir_i}) \tag{8}$$

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 $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}z}$. 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 [21], 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 [22]

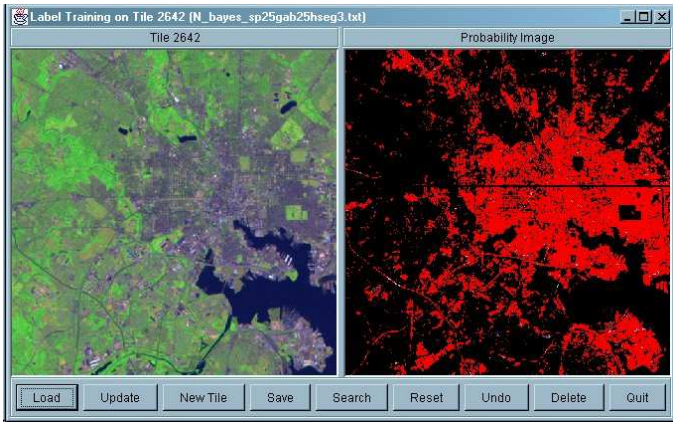


Fig. 3. Training for the city prototype. Positive and negative examples of city pixels in the image on the left are used to learn a Bayesian classifier that creates the probability map shown on the right. Brighter values in the map show pixels with high probability of being part of a city. Pixels marked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

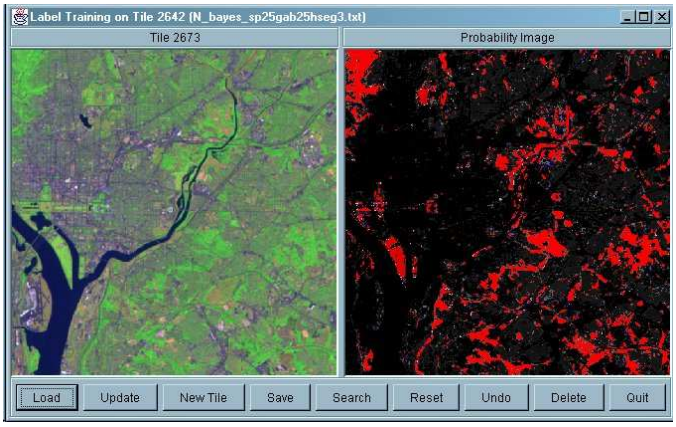


Fig. 4. Training for the park prototype using the process described in Fig. 3.

to automatically divide large regions into more compact sub-regions. 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,
2) Mark regions with areas smaller than a threshold as background using connected components analysis [22],
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 [22] and repeat:
   a) Threshold erosion transform at steps of 3 pixels in every iteration,
   b) Find connected components of the thresholded image,
   c) Select sub-regions that have an area smaller than a threshold,
   d) Dilate these sub-regions to restore the effects of erosion,
   e) Mark these sub-regions in the output image by masking the dilation using the original image,

   until no more sub-regions are found,
6) Merge the residues of previous iterations to their smallest neighbors.

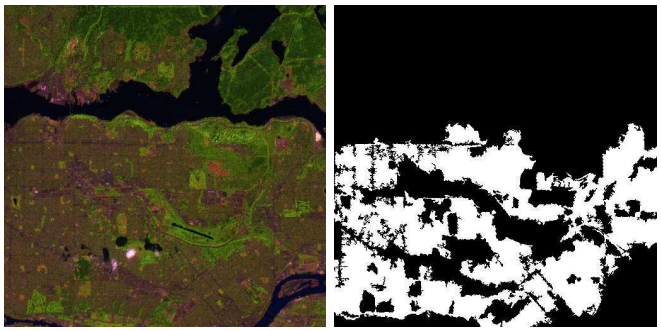
The merging and splitting process is illustrated in Fig. 5. The probability of each region belonging to a land cover label can be estimated by propagating class labels from pixels to regions. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the set of pixels that are merged to form a region. Let w_j and $p(w_j|x_i)$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

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

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

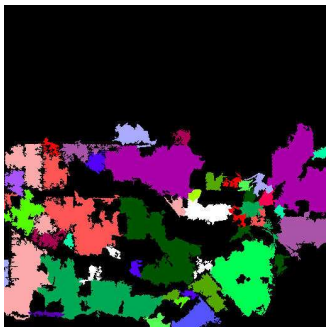
Fig. 6 shows example segmentations. The number of clusters in k-means clustering was empirically chosen as 25 both for spectral values and for Gabor features. The number of clusters for hierarchical segmentation features was automatically obtained as 17. The probability threshold and the minimum area 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, micro-texture 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, hyper-spectral, 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 is available for the same area, a single Bayes classifier does



(a) LANDSAT image

(b) A large connected region formed by merging pixels labeled as residential



(c) More compact sub-regions

Fig. 5. Region segmentation process. The iterative algorithm that uses mathematical morphology operators is used to split a large connected region into more compact sub-regions.

automatic fusion for a particular label as given in (1) and described above. If different sensors are available for different areas in the same data set, 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.

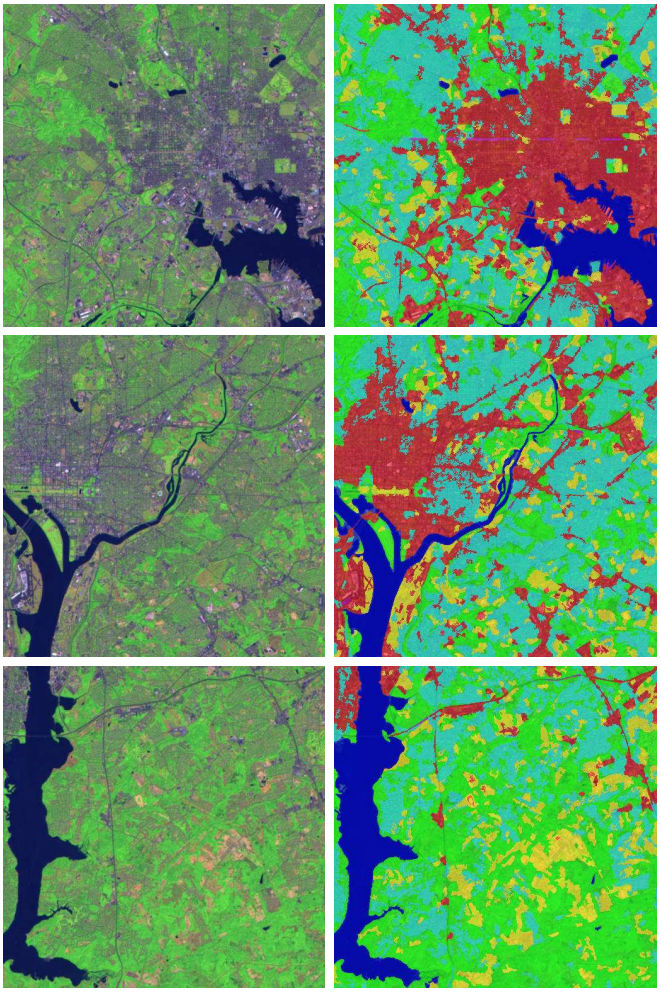


Fig. 6. Segmentation examples from the NASA data set. Images on the left column are used to train pixel level classifiers for city, residential area, water, park and field using positive and negative examples for each class. Then, these pixels are combined into regions using the iterative region split-and-merge 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.

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.

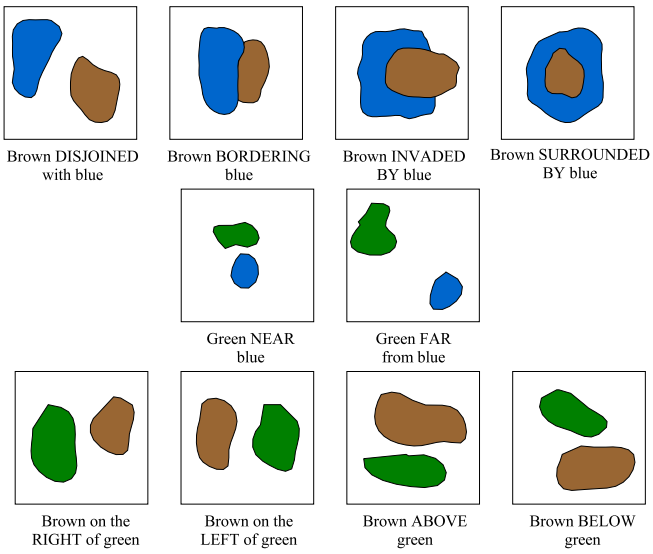


Fig. 7. Spatial relationships of region pairs: *disjoined*, *bordering*, *invaded_by*, *surrounded_by*, *near*, *far*, *right*, *left*, *above* and *below*.

– *below*: First region is below the second one.

These relationships are illustrated in Fig. 7. They are divided into sub-groups 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

- perimeter of the first region, π_i
- perimeter of the second region, π_j
- common perimeter between two regions, π_{ij}, computed as the shared boundary between two polygons
- ratio of the common perimeter to the perimeter of the first region, $r_{ij} = \frac{\pi_{ij}}{\pi_i}$
- 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, ν_i
- centroid of the second region, ν_j
- angle between the horizontal (column) axis and the line joining the centroids, θ_{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 θ_{ij} with truncated cosine membership functions. Details of the membership functions are given in [12]. 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 [22] 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 labeled with the spatial relationship classes (pairwise relationship names) and the corresponding degrees (fuzzy membership values) for these relationships.

V. Image Classification

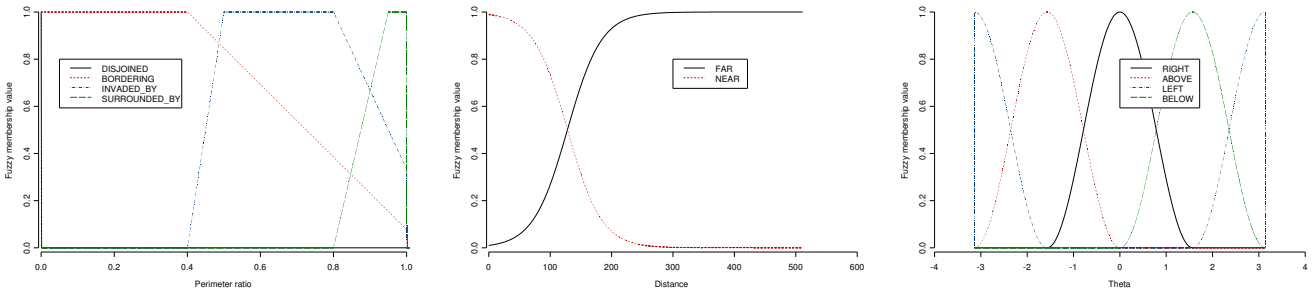
Image classification is defined here as a problem of assigning images to different classes according to the scenes they contain (the last process in Fig. 1). The visual grammar enables creation of high-level classes that cannot be modeled by individual pixels or regions. Furthermore, learning of these 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 $\sigma_W^2 = \sum_{i=1}^{s} v_i \text{var}\{z_j \,|\, j \in w_i\}$ is the within-class variance, v_i is the number of training images for



(a) Perimeter-class relationships (b) Distance-class relationships (c) Orientation-class relationships

Fig. 8. Fuzzy membership functions for pairwise spatial relationships.

class w_i, z_j is the number of times this region group is found in training image j, $\sigma_B^2 = \text{var}\{\sum_{j \in w_i} z_j \,|\, i = 1, \ldots, s\}$ is the between-class variance, and 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 Binomial$(v_i, \theta_j) = \binom{v_i}{v_{ij}} \theta_j^{v_{ij}} (1 - \theta_j)^{v_i - v_{ij}}$ 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 θ_j becomes

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

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 MAP rule with the conditional independence assumption as

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

Classification examples from the PRISM data set 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.

[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.

VI. Conclusions

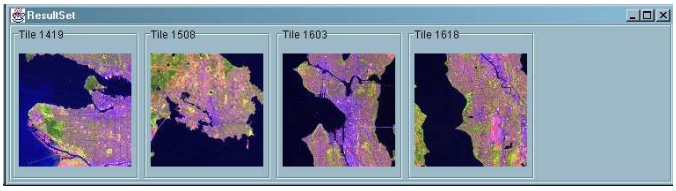
We described a visual grammar that aims to bridge the gap between low-level features and high-level semantic interpretation of images. The system uses naive Bayes classifiers to learn models for region segmentation and classification from automatic fusion of features, fuzzy modeling of region spatial 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

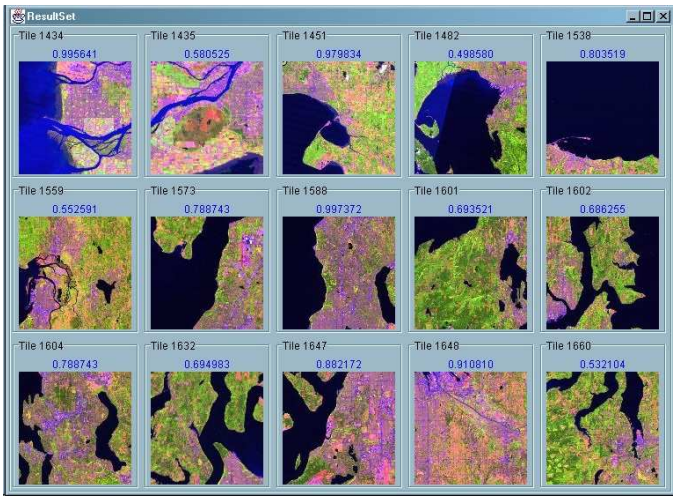
(a) Training images for clouds

(b) Images classified as containing clouds

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 residential areas with a coastline

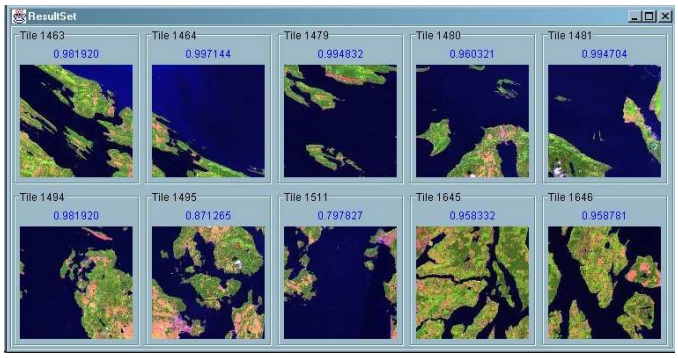


(b) Images classified as containing residential areas with a coastline

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).

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.

(a) Training images for tree covered islands



(b) Images classified as containing tree covered islands

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).

REFERENCES

[1] NASA, "TERRA: The EOS flagship," http://terra.nasa.gov.

[2] M. Schroder, H. Rehrauer, K. Siedel, and M. Datcu, "Interactive learning and probabilistic retrieval in remote sensing image archives," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 5, pp. 2288–2298, September 2000.

[3] S. I. Hay, M. F. Myers, N. Maynard, and D. J. Rogers, Eds., *Photogrammetric Engineering & Remote Sensing*, vol. 68, no. 2, February 2002.

[4] K. Koperski, G. Marchisio, S. Aksoy, and C. Tusk, "VisiMine: Interactive mining in image databases," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, vol. 3, Toronto, Canada, June 2002, pp. 1810–1812.

[5] A. C. Shaw, "Parsing of graph-representable pictures," *Journal of the Assoc. of Computing Machinery*, vol. 17, no. 3, pp. 453–481, July 1970.

[6] J. R. Smith and S.-F. Chang, "VisualSEEk: A fully automated content-based image query system," in *Proceedings of ACM International Conference on Multimedia*, Boston, MA, November 1996, pp. 87–98.

[7] P. J. Neal, L. G. Shapiro, and C. Rosse, "The digital anatomist structural abstraction: A scheme for the spatial description of anatomical entities," in *Proceedings of American Medical Informatics Association Annual Symposium*, Lake Buena Vista, FL, November 1998.

[8] W. W. Chu, C.-C. Hsu, A. F. Cardenas, and R. K. Taira, "Knowledge-based image retrieval with spatial and temporal constructs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 10, no. 6, pp. 872–888, November/December 1998.

[9] 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 *Proceedings of SPIE Medical Imaging*, vol. 3662, San Diego, CA, February 1999, pp. 360–368.

[10] E. G. M. Petrakis and C. Faloutsos, "Similarity searching in medical image databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 3, pp. 435–447, May/June 1997.

[11] S. Aksoy, G. Marchisio, K. Koperski, and C. Tusk, "Probabilistic retrieval with a visual grammar," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, Toronto, Canada, June 2002, pp. 1041–1043.

[12] 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. World Scientific, 2003, pp. 35–62.
[13] G. M. Haley and B. S. Manjunath, "Rotation-invariant texture classification using a complete space-frequency model," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 255–269, February 1999.
[14] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, November 1973.
[15] J. C. Tilton, G. Marchisio, K. Koperski, and M. Datcu, "Image information mining utilizing hierarchical segmentation," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, Toronto, Canada, June 2002, pp. 1029–1031.
[16] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
[17] M. H. DeGroot, *Optimal Statistical Decisions*. McGraw-Hill, 1970.
[18] D. Geiger and D. Heckerman, "A characterization of the Dirichlet distribution through global and local parameter independence," *The Annals of Statistics*, vol. 25, no. 3, pp. 1344–1369, 1997.
[19] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.
[20] R. F. Krichevskiy, "Laplace's law of succession and universal encoding," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 296–303, January 1998.
[21] S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, and J. C. Tilton, "Learning Bayesian classifiers for a visual grammar," in *Proceedings of IEEE GRSS Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, Washington, DC, October 2003, pp. 212–218.
[22] R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*. Addison-Wesley, 1992.

Selim Aksoy (S'96-M'01) received his B.S. degree from Middle East Technical University in 1996, and his 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 at the Department of Computer Engineering at Bilkent University, Turkey. Before joining Bilkent, he was a Research Scientist at Insightful Corporation in Seattle, where he was involved in image understanding and data mining research sponsored by NASA, U.S. Army and 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, relevance feedback and mathematical morphology. During summers of 1998 and 1999, he was a Visiting Researcher at the Tampere International Center for Signal Processing, Finland, 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. He is a member of IEEE and the International Association for Pattern Recognition (IAPR). Dr. Aksoy was recently elected as 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, British Columbia, in 1999. During his graduate work at Simon Fraser University, Krzysztof Koperski worked on knowledge discovery in spatial databases and spatial data warehousing. In 1999 he was a Visiting Researcher at University of L'Aquila, Italy working on spatial data mining in the presence of uncertain information. Since 1999 Dr. Koperski has been with Insightful Corporation based in Seattle, Washington. 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 (RWTH) in Aachen, Germany in 2001. Since August 2001 he is working as a Research Scientist at Insightful Corporation, Seattle, USA. His research interests are in information retrieval, statistical data analysis and database systems.

Giovanni Marchisio is currently Director Emerging Products at Insightful Corporation, Seattle. 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 PI on R&D government contracts totaling several millions (with NASA, NIH, DARPA, and DoD). 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. Dr. Marchisio has authored and co-authored 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 US patents. Dr. Marchisio has also been a visiting professor at the University of British Columbia, Vancouver, Canada. His previous work includes research in signal processing, ultrasound acoustic imaging, seismic reflection and electromagnetic induction imaging of the earth interior. Giovanni holds a B.A.Sc. in Engineering from University of British Columbia, Vancouver, Canada, and a Ph.D. in Geophysics and Planetary Physics from University of California San Diego (Scripps Institute).

James C. Tilton (S'79-M'81-SM'94) received B.A. degrees in electronic engineering, environmental science and engineering, and anthropology and a M.E.E. (electrical engineering) from Rice University, Houston, TX in 1976. He also received an M.S. in optical sciences from the University of Arizona, Tucson, AZ in 1978 and a Ph.D. in electrical engineering from Purdue University, West Lafayette, IN in 1981. He is currently a Computer Engineer with the Applied Information Science Branch (AISB) of the Earth and Space Data Computing Division at the Goddard Space Flight Center, NASA, Greenbelt, MD. He previously worked for Computer Sciences Corporation from 1982 to 1983 and Science Applications Research from 1983 to 1985 on contracts with NASA Goddard. As a member of the AISB, Dr. Tilton is responsible for designing and developing computer software tools for space and earth science image analysis, 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. Dr. Tilton is a senior 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, and since 2001 he has served as an Associate Editor for the Pattern Recognition journal.