# Learning Spatial Configuration Models Using Modified Dirichlet Priors

**Matthew Boutell** BOUTELL@CS.ROCHESTER.EDU
**Christopher Brown** BROWN@CS.ROCHESTER.EDU
Department of Computer Science, University of Rochester, Rochester, NY 14627 USA

**Jiebo Luo** JIEBO.LUO@KODAK.COM
Research and Development Labs, Eastman Kodak Company, Dewey Ave., Rochester, NY 14650 USA

## Abstract

Semantic scene classification is a challenging problem in computer vision. Special-purpose semantic object and material (e.g., sky and grass) detectors help, but are faulty in practice. In this paper, we propose a generative model of outdoor scenes based on spatial configurations of objects in the scene. Because the number of semantically-meaningful regions (for classification purposes) in the image is expected to be small, we infer exact probabilities by utilizing a brute-force approach. However, it is impractical to obtain enough training data to learn the joint distribution of the configuration space.

To help overcome this problem, we propose a smoothing technique that modifies the naive uniform (Dirichlet) prior by using model-based graph-matching techniques to populate the configuration space. The proposed technique is inspired by the backoff technique from statistical language models. We compare scene classification performance using our method with two baselines: no smoothing and smoothing with a uniform prior. Initial results on a small set of natural images show the potential of the method. Detailed exploration of the behavior of the method on this set may lead to future improvements.

## 1. Introduction

Semantic scene classification, categorizing photographs at a high level into discrete categories such as *beach*, *mountain*, or *indoor*, is a useful, yet challenging problem in computer vision. It can help with image organization and with content-based image retrieval. Most approaches (Vailaya et al., 1999; Szummer & Picard, 1998; Torralba et al., 2003) typically use low-level (e.g., color, texture) features and classifiers to achieve reasonable results.

Higher-level features, such as the output from object and material detectors, can also help classify scenes. An advantage to this approach is its modularity, allowing independently-developed, domain-sensitive detectors to be used. Only recently has object and material detection in natural environments become accurate enough to consider using in a practical system. Recent work using object detection for other tasks (Mulhem et al., 2001; Song & Zhang, 2002; Smith et al., 2003) has achieved some success using object presence or absence alone as evidence. However, faulty detectors present a continuing difficulty for this approach.

Figure 1 shows an image, true material identities of key regions (color-coded), and simulated detector results, expressed as likelihoods that each region is labeled with a given material. The problem is how to determine which scene type best explains the observed (imperfect) evidence.

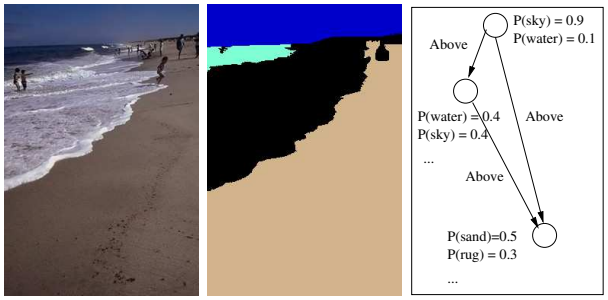


*Figure 1.* (a) A beach image (b) Its manually-labeled materials. The true configuration includes *sky above water*, *water above sand*, and *sky above sand*. (c) The underlying graph showing detector results and spatial relations.

How does one overcome detector errors? One principled way is to use a probabilistic inference system (vs. a rule-based one, such as (Mulhem et al., 2001)) to classify a scene based on the presence or absence of semantic materials. Furthermore, we can extract additional useful evidence from the input image, such as spatial relationships between the detected regions, to improve scene classification.

In this paper, we study statistical relational learning of *scene configurations*, consisting of both materials *and* their spatial relations. We propose a generative model of scene classification that uses material detectors and full scene configurations. The main limitation of this model is obtaining enough training data to learn the joint distribution of the *configuration space* (materials in specific configurations). To this end, we also propose a smoothing technique that improves upon the naive uniform (Dirichlet) prior by using model-based graph-matching techniques to populate the configuration space. Our technique is inspired by graph matching and backoff techniques. It is used in learning only; the inference phase needs no adaptation. We compare scene classification performance using our method with two baselines, no smoothing and a uniform prior, to show its promise.

## 2. Scene Configurations

Scene configurations consist of both the actual spatial arrangement (graph edge labels) of regions and the identities of those regions (node labels), as illustrated in Figure 1.

Our terminology is as follows: let $n$ be the number of distinct regions detected in the image, $M$ be the small set of semantically critical materials for which detectors are used, $SR$ be the set of spatial relations, and $C$ be the set of configurations of materials in a scene. Then an upper bound on the number of scene configurations, $|C|$, is $|M|^n|SR|^{\binom{n}{2}}$ (in a fully connected graph).

In (Luo et al., 2003), the spatial relations *above, below, far above, far below, beside, enclosed*, and *enclosing* (i.e., $|SR| = 7$)) were shown to be effective for spatially-aware material detection within outdoor scenes. We adopt essentially the same spatial relations in this study.

In the inference phase, the spatial arrangement of the test image is known (computed); thus, its graph need only be compared with those of training images with the same arrangement. We restrict our attention in this paper to learning the distribution of region identities within a *fixed* spatial arrangement, of which there are $|M|^n$ configurations. For example, an image with two regions, $R_1$ *above* $R_2$ has only $|M|^2$ configurations.

Adding to our terminology, we can formalize the scene classification problem as follows: let $S$ be the set of scene classes considered, and $E = \{E_1, E_2, ...E_n\}$ be the detector evidence, where each $E_j = \{e_1, e_2, \dots e_{|M|}\}$ is a likelihood vector for the identity of region $j$.

In this framework, we seek to calculate:

$$\arg\max_i P(S_i|E) \propto \arg\max_i P(S_i)P(E|S_i) \quad (1)$$

Taking the joint distribution of $P(E|S_i)$ with $C$ yields

$$\arg\max_i P(S_i)\sum_{c\in C} P(E, c|S_i) \quad (2)$$

Conditioning on $c$ gives

$$\arg\max_i P(S_i)\sum_{c\in C} P(E|c, S_i)P(c|S_i) \quad (3)$$

### 2.1. Relation to Graphical Models

Figure 2a shows a graphical *representation* (not graphical model) for a single scene. While it is possible to represent this system with a grapical model, we chose a different approach in this study. On the surface, it looks like a standard two-level Markov Random Field (MRF) (Geman & Geman, 1984; Freeman et al., 2000). As in these MRFs, evidence nodes in our representation are conditionally dependent only on the identity of the underlying region's scene node, while the scene nodes are dependent on each other. However, this is not a typical graphical model. Fundamentally, we are solving a different problem than those for which MRFs are used. MRFs are typically used to regularize input data (Geman & Geman, 1984; Chou, 1988), finding $P(C|E)$, the single configuration (within a single model) that best explains the observed faulty evidence. In constrast, we are trying to perform *cross-model comparison*, $P(S|E)$, comparing how well the evidence matches each model in turn. To do this, we need to sum across *all possible* configurations of the scene nodes (Equation 3). In this framework, we need to use a brute force approach.

Further, at this coarse segmentation, even distant (in the underlying image) nodes may be strongly correlated, e.g., sky and pavement in urban scenes. Thus, we cannot factorize the scene structure (as could be done in low-level vision problems) and instead assume a fully-connected scene structure. Fortunately, for

scene classification, and particularly for landscape images, the number of critical material regions of interest, $n$, is generally small ($n \leq 6$) [1], so a brute-force approach to maximizing Equation 3 is tractable.

### 2.2. Scene Classification System

Figure 2b shows the full scene classification system. After each scene model computes $\sum_{c \in C} P(E|c, S_i)P(c|S_i)$ for scene class $S_i$, the results are compared at the top level to make a decision (incorporating priors for the scene classes if available).

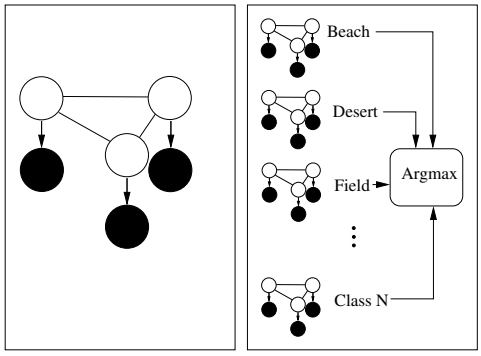


*Figure 2.* (a) Graphical representation for a single scene class. The observed nodes (detector inputs; shown as filled circles) are each connected to a single node in the fully-connected scene graph (which represents a configuration, treated as a single hidden state in the brute-force approach). (b) Full system: a bank of scene models. The instantiated detector inputs to each scene are the same for each class.

We can learn $P(E|c, S)$ relatively easily. As described above, a reasonable assumption is that a detector's output on a region depends only on the object present in that region and not on other objects or upon the class of the scene. This allows us to factor the distribution into $\prod_{j=1}^{n} P(E_j|c_j)$; each of which describes a single detector's characteristics and can be learned by counting detection frequencies on a training set of regions or fixed using using domain knowledge.

However, $P(c|S)$ is difficult to factor because of the strong dependency between regions. The resulting joint distribution is sparsely populated: there are $O(|M|^n)$ parameters (the counts of each configuration) to learn, and only $|T_S|$ training images of scene class $S$. The sparseness is exacerbated by correlation between objects and scenes. How do we deal with this sparse distribution?

[1]The material detectors can be imbued with the ability to merge regions, so over-segmentation is rare.

### 2.3. Naive Approaches to Smoothing

The simplest approach is to do nothing. This adds no ambiguity to the distribution. However, without smoothing, we have $P(c|S) = 0$ for each configuration $C \notin T_S$. This automatically rules out, by giving zero probability to, any valid configuration not seen in the sparse training set, regardless of the evidence: clean, but unsatisfying.

Another simple technique is use a uniform Dirichlet prior on the configurations, implemented by adding a matrix of pseudo-counts of value $\epsilon$ to the matrix of configuration counts. However, this can allow for too much ambiguity, because in practice, many configurations should be impossible, for example configurations containing *snow* in the *Desert* model. We seek the middle ground: allowing some matches with configurations not in the training set, but not indiscriminately allowing all matches.

### 2.4. Graph-based Smoothing

Our goal is to smooth using the training set and knowledge of the image domain. Specifically, we compute $P(c|S)$ as follows. Fix the spatial arrangement of materials. Let $T_S = \{G_{1,S}, G_{2,S}, \ldots G_{|T_S|,S}\}$ be the set of graphs of training images of class $S$ with that spatial arragnment. For $1 \leq j \leq r$, let $N_S^j \in M^r$ be $r$-dimensional matrices of counts. The configuration, $c$, is an index into the matrices. Let $sub_j(G)$ denote a subgraph of graph $G$ with $j$ nodes and $\equiv$ denote graph isomorphism.

Then define

$$N_S^j(c) = |\{G_{i,S}\}| : sub_j(c) \equiv sub_j(G_{i,S})) \quad (4)$$

$$N_S(c) = \sum_{j=1}^{r} \alpha_j N_S^j(c) \quad (5)$$

$$P(c|S) = \frac{N_S(c)}{\sum_{\tilde{c} \in C} N_S(\tilde{c})} \quad (6)$$

Each $N^j$ represents the pseudo-counts of the subgraphs of size $j$ occurring in the training set. $N^r$ is the standard count of full scene configurations occurring in the training set. As the subgraph size decreases, the subgraphs are less-specific to the training set, and so should contribute less to the overall distribution. Thus, we expect that the parameters $\alpha_j$ will decrease monotonically. Furthermore, as $j$ decreases, each $N_j$ becomes more densely populated. Intuitively, this is like radial basis function smoothing, in that points

"close" to the training points are given more weight in the small area near the peaks than in the larger area at the tails. Finally the counts are normalized to obtain $P(c|S)$. For example, consider the contribution to $N$ of a *single* training configuration "sky above water above sand": each configuration containing "sky above sand", "sky above water", or "sand above water" receives weight $\alpha_2$ and any configuration containing sky, water, or sand receives weight $\alpha_1 < \alpha_2$; other configurations receive no weight.

The desired result of modifying the uniform Dirichlet prior in this way is that the weight a configuration receives is a function of how closely it matches configurations seen in the training set. While our proposed technique is inspired mainly by the graph matching literature, it can also be viewed as backprojection and as a backoff technique; we discuss each connection in Section 4.

## 3. Experimental Results

We have a database of 923 images in 5 classes: Beach, Desert, Fall Foliage, Field, and Mountain. Each image is automatically segmented using the algorithm described in (Comaniciu & Meer, 2002), and the semantically-critical regions are manually labeled with their true materials (i.e., ground truth), as in Figure 1b. The ground truth labels correspond to those materials (e.g., sky, grass, foliage, rocks) expected to predict these scenes. Other regions are left unlabeled.

To simulate actual detectors, which are faulty, we randomly perturbed the ground truth to create simulated detector responses. We set the detection rates of individual material detectors on each true material (both *true positive rates*, e.g., how often the grass detector fires on grass regions, and *false positive rates*, e.g., how often the grass detector fires on water regions) by counting performance of corresponding actual detectors on a validation set (or estimating them in the case of detectors to be developed in the future). When they fire, they are assigned a belief that is distributed normally with mean $\mu$. The parameter $\mu$ can be set differently for true and false positive detections; varying the ratio between the two is a convenient way to simulate detectors with different accuracies.

Spatial relations are computed using a computationally efficient version of the "weighted walk-through" approach (Berretti et al., 2001), detailed in (Luo et al., 2003). We simplify the relations by ignoring the "far" modifier and the enclosure relations (which occur rarely). We focus further on the single spatial arrangement occurring most often in training: of the 256 images with 3 regions, 172 have a vertical structure, $R_1$ *above* $R_2$, $R_2$ *above* $R_3$, and $R_1$ *above* $R_3$.

We learn $P(c|S)$ using the proposed smoothing method and compare it with learning using two baselines: no smoothing and smoothing by a uniform prior added to the counts.

We perform leave-one-out cross-validation to estimate scene classification performance using the brute force inference method of Equation 3. We eliminate the effect of the priors over scene types by setting them equal. Because we are simulating faulty detectors, we can vary their performance and compare the methods across the spectrum of detector accuracy (Figure 3). While the subgraph smoothing method performs better than the two baselines at all detector accuracies, we admit the difference is small. We believe optimizing the smoothing weights, $\alpha_j$, should improve performance; learning those parameters is the subject of future work.

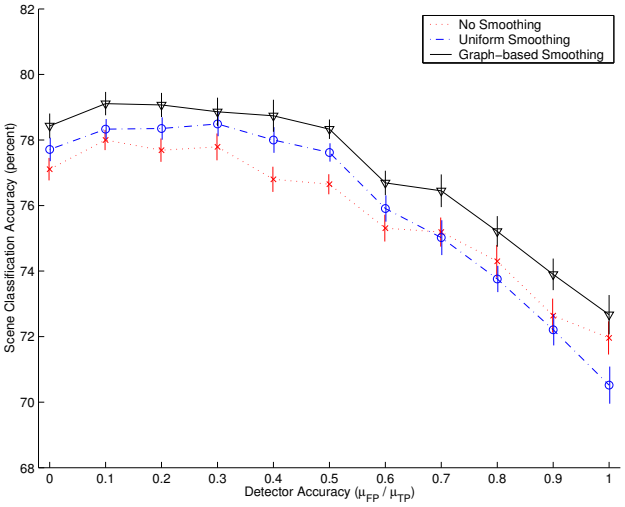


*Figure 3.* Classification accuracy of the methods as a function of detector accuracy. The smoothing method performs better than the baselines at nearly all detector accuracies. Standard error is shown ($n = 30$).

## 4. Discussion

As expected, the smoothing technique helps to classify images having a plausible, but rarely-occurring, scene configuration. Figure 4 shows a number of examples. The mountain scene on the left with the configuration "gray sky above snow above grass" was classified correctly by our method, but failed when no smoothing was applied, because that specific configuration did not occur in the training set, but its subgraphs did. Other similar examples are the desert scene in

the configuration "blue sky above blue sky above bare ground" and the field scene in the configuration "foliage above blue sky above grass".

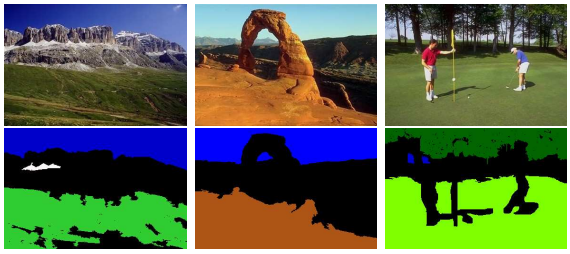

*Figure 4.* Some images for which the baseline methods fail, but the proposed method succeeds. Top: original scenes. Bottom: hand-labeled regions.

### 4.1. Related Techniques for Graph Matching

Presently we are doing *exact* graph matching in the sense that we demand an isomorphism for the arcs and nodes, but *inexact* matching in that we are matching *attributed* graphs, those with values or labels attached to the nodes and arcs. We do *multiple-matching*: we are matching into a database of graphs, looking for the best match. Graph isomorphism is a problem of unknown complexity. Inexact graph matching (differing number of nodes) is known to be NP-complete, but is a basic utility for recognition problems. Thus graph matching has a long history in image and scene understanding.

Probabilistic techniques in graph matching, often using relaxation, have been used for some time (Hancock & Kittler, 1990). A comparison of various search strategies appears in (Williams et al., 1999), and (Shams et al., 2001) compares various matching algorithms to one based on maximizing mutual information. Hierarchical relaxation has been used in a Bayesian context (Wilson & Hancock, 1999). Mixture models have been explored for EM matching: a weighted sum of Hamming distances is used as a matching metric in (Finch et al., 1998). Generally, only unary attributes and binary relations are used in these probabilistically-founded searches. More complex relations can be used in relaxation-like schemes as in (Skomorowski, 1999). Various schemes using learning and Bayes nets for inexact matching are explored in (Bengoetxea et al., 2000).

### 4.2. Related Concepts

One way to view our method is as a backprojection (Swain & Ballard, 1991) technique. If we view the configuration space $C$ as an $r$-dimensional space, subgraphs of lower dimension (size $i < r$) can be backprojected into $C$ to populate the space. Figure 5 shows an example with $r = 3$ and $i = 2$. The 3D configuration space is sparsely populated in the absence of smoothing. The training points are projected into 2D along 3 axes (the three subgraphs). The resulting 2D spaces corresponding to the same spatial configurations are integrated to combine evidence from different training examples. Finally they are backprojected into the original 3D space (with lower weights than the original counts).

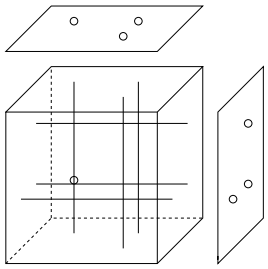

*Figure 5.* Backprojection using our technique. For legibility in this 3D example, only one training point and two backprojection directions (of the three possible with this spatial configuration) are shown. The training set generalizes through combining counts of subgraphs of multiple training configurations.

Our technique can also be viewed as a backoff technique, as commonly used in speech recognition (Manning & Schutze, 1999); a hierarchical, more principled model is presented in (MacKay & Peto, 1994). If there is insufficient data to learn a trigram model for a given word, one can use a less-specialized, but more densely-populated, bigram or unigram model. However, we pre-compute the probabilities in the learning phase; inference needs no special treatment.

## 5. Conclusions and Future Work

We have presented a generative model for classifying scenes using faulty material detectors and spatial configurations of materials present in the image. This approach poses a challenge to statistical relational learning, as scene configurations attempt to capture correlations between sets of materials and scene types. Initial results on a small set of landscape images have also shown that performance can be improved by using a smart smoothing technique using subgraphs of the training images.

Clearly, this is work in progress. Future investigation will involve experimentation using real material detectors and a much larger number of images. We also plan to expand the library of spatial arrangements (e.g., $R_1$ above $R_2$, $R_1$ above $R_3$, $R_2$ *beside* $R_3$) and to address

the accompanying scalability issues through investigating prototypical spatial arrangements and factorization of the scene models.

More detailed analysis of the behavior of the method may lead to future improvements, such as in learning the parameters of the model. Another interesting direction is to incorporate theoretical work on improving purely uniform priors (Nemenman et al., 2001).

## Acknowledgments

This research was supported by a grant from Eastman Kodak Company and by the NSF under grant number EIA-0080124.

## References

Bengoetxea, E., Larranaga, P., Bloch, I., Perchanta, A., & Boeres, C. (2000). Inexact graph matching using learning and simulation of Bayesian networks. *Proc. CaNew workshop, ECAI 2000*.

Berretti, S., Bimbo, A. D., & Vicario, E. (2001). Spatial arrangement of color in retrieval by visual similarity. *Pattern Recognition*, *35*, 1661–1674.

Chou, P. (1988). *The theory and practice of Bayesian image labeling*. Doctoral dissertation, University of Rochester, Rochester, NY.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 603–619.

Finch, A. W., Wilson, R. C., & Hancock, E. R. (1998). Symbolic graph matching with the EM algorithm. *Pattern Recognition*, *31*, 1777–1790.

Freeman, W., Pasztor, E., & Carmichael, O. (2000). Learning low-level vision. *International Journal of Computer Vision*, *40*, 24–57.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Hancock, E. R., & Kittler, J. (1990). Edge-labeling using dictionary-based relaxation. *IEEE-TPAMI*, *12*, 165–181.

Luo, J., Singhal, A., & Zhu, W. (2003). Towards holistic scene content classification using spatial context-aware scene models. *IEEE Conference on Computer Vision and Pattern Recognition*. Madison, WI.

MacKay, D., & Peto, L. (1994). A hierachical Dirichlet language model. *Natural Language Engineering*, *1*, 1–19.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mulhem, P., Leow, W. K., & Lee, Y. K. (2001). Fuzzy conceptual graphs for matching images of natural scenes. *IJCAI* (pp. 1397–1404).

Nemenman, I., Shafee, F., & Bialek, W. (2001). Entropy and inference, revisited. *NIPS*.

Shams, L., Brady, M., & Schall, S. (2001). Graph matching vs. mutual information maximization for object detection. *Neural Networks*, *14*, 345–354.

Skomorowski, M. (1999). Use of random graph parsing for scene labelling by probabilistic relaxation. *Pattern Recognition Letters*, *60*, 949–956.

Smith, J., Lin, C., Naphade, M., Natsev, A., & Tseng, B. (2003). Multimedia semantic indexing using model vectors. *IEEE ICME*. Baltimore, MD.

Song, Y., & Zhang, A. (2002). Analyzing scenery images by monotonic tree. *ACM Multimedia Systems Journal*.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*.

Szummer, M., & Picard, R. W. (1998). Indoor-outdoor image classification. *IEEE International Workshop on Content-based Access of Image and Video Databases*. Bombay, India.

Torralba, A., Murphy, K., Freeman, W., & Rubin, M. (2003). Context-based vision system for place and object recognition. *International Conference on Computer Vision*.

Vailaya, A., Figueiredo, M., Jain, A., & Zhang, H. (1999). Content-based hierarchical classification of vacation images. *Proc. IEEE Multimedia Systems '99 (International Conference on Multimedia Computing and Systems)*. Florence, Italy.

Williams, M. L., Wilson, R. C., & Hancock, E. R. (1999). Deterministic search for relational graph matching. *Pattern Recognition*, *32*, 1255–1271.

Wilson, R. C., & Hancock, E. R. (1999). Graph matching with hierarchical discrete relaxation. *Pattern Recognition Letters*, *20*, 80–96.