Multiscale Conditional Random Fields for Image Labeling

Xuming He   Richard S. Zemel   Miguel Á. Carreira-Perpiñán
Department of Computer Science, University of Toronto
{*hexm,zemel,miguel*}*@cs.toronto.edu*

Abstract

We propose an approach to include contextual features for labeling images, in which each pixel is assigned to one of a finite set of labels. The features are incorporated into a probabilistic framework which combines the outputs of several components. Components differ in the information they encode. Some focus on the image-label mapping, while others focus solely on patterns within the label field. Components also differ in their scale, as some focus on fine-resolution patterns while others on coarser, more global structure. A supervised version of the contrastive divergence algorithm is applied to learn these features from labeled image data. We demonstrate performance on two real-world image databases and compare it to a classifier and a Markov random field.

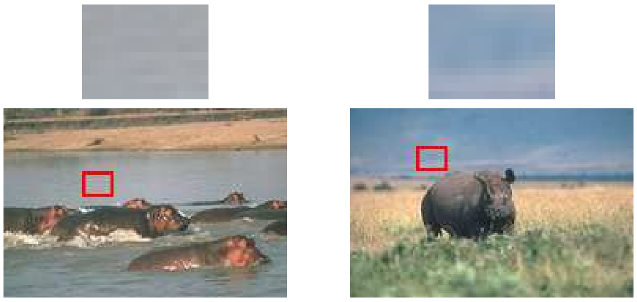

Figure 1: Top: two small image patches that are difficult to label based on local information. Bottom: images containing the patches. The global context makes it clear what the patches are (left: water; right: sky).

1. Introduction

We consider the following problem, that we will call *image labeling*: to classify every pixel of a given image into one of several predefined classes. For example, we might consider images of wildlife in the savanna, and we would like to classify each pixel as either animal, ground, vegetation, water or sky. The result is both a segmentation of the image and a recognition of each segment as a given object class. Automatically labeled images can also be useful for other purposes, such as for image database querying (e.g. "find all images with animals in the water"). In this paper, we propose a model that is learned from labeled images.

Labeling requires *contextual* information, because the labels are dependent across pixels. Further, an image contains information that is useful for labeling at several levels. At a local level (a few pixels wide), the color and texture can sometimes be enough to identify the pixel class—e.g. the sky tends to be uniformly blue. However, typically this is complicated by the large overlap between classes (the water can also be blue) and the noise in the image. An example is given in Figure 1: two small image patches are ambiguous at a very local scale but clearly identifiable inside their context. Aspects of this context concern the geometric relationships between objects—e.g. fish tend to be in water and airplanes in the sky; while other aspects concern the location of objects in the image—e.g. the sky tends to be at the top of the image and the water at the bottom. Thus, context at a level more than a few pixels wide can help disambiguate very local image information.

We thus have information about the labeling coming from different scales (local and global). This presents two problems: First, how can we extract and represent the information at each level? Second, how should we combine the possibly conflicting information from the different levels?

1.1. Previous Approaches

One response to these questions is offered by a common approach to region classification, Markov random fields (MRFs). MRFs are typically formulated in a probabilistic generative framework modeling the joint probability of the image and its corresponding labels [6, 12]. MRFs suffer from two key limitations with respect to the labeling problem. The first drawback concerns their locality. Generally, due to the complexity of inference and parameter estimation, only local relationships between neighboring nodes are incorporated into the model. This allows the model to locally smooth the assigned labels, based on very local regularities, but makes it highly inefficient at capturing long-range interactions. However, as discussed above, the conditional probability of a labeling will likely depend on structure at different levels of granularity in the image. We seek a model that can capture both local as well as more global relationships. Hierarchical MRFs [1, 10] offer one way of capturing label relationships at different scales, but still suffer from the second main drawback of MRFs, which lies in their generative nature. Many labeled images are required to estimate the parameters of the model of labels *and* images. We are interested in estimating the posterior over labels given the observed image; even when this posterior is simple, the true underlying generative model may be quite complex. Because we are only interested in the distribution of labels given images, devoting model resources and degrees-of-freedom to the generative image model is unnecessary.

A very different *non-generative* approach is to directly model the conditional probability of labels given images: fewer labeled images will be required, and the resources will be directly relevant to the task of inferring labels. This is the key idea underlying the conditional random field (CRF) [11]. Originally proposed for segmenting and labeling 1-D text sequences, CRFs directly model the posterior distribution as a Gibbs field. This conditional probability model can depend on arbitrary non-independent characteristics of the observation, unlike a generative image model which is forced to account for dependencies in the image, and therefore requires strict independence assumptions to make inference tractable. CRFs have been shown to outperform traditional hidden Markov model labeling of text sequences [11].

In this paper, we aim to generalize the CRF approach to the image labeling problem, which is considerably more complicated due to the 2-D nature of images versus the 1-D nature of text. We also aim to learn features in the random field that operate at different scales of the image. We adopt a statistical learning approach, where such information is learned from a training set of labeled images, and combined in a probabilistic manner.

2. Multiscale Conditional Random Field

Let $\mathbf{X} = \{\mathbf{x}_i\}_{i \in S}$ be the observed data from an input image where S is a set of image sites to be labeled. We use the term sites to refer to elements of the label field, while pixels refer to elements of the image. The local observation $\mathbf{x}_i$ at site i is the response of a set of filters applied to the image at that site. The site has an associated label l_i from a finite label set $\mathcal{L}$.

Standard CRFs [11] employ two forms of feature functions, which would be defined in a 2D image as follows: state feature functions, $f(l_i, \mathbf{X}, i)$, of the label at a site i and the observed image; and transition feature functions $f(l_i, l_j, \mathbf{X}, i)$, of the image and labels at site i and a neighboring site j in the image. We extend this to *label features*, which encode particular patterns within a subset of label variables. The label features are a form of potential function, encoding a particular constraint between the image and the labels within a region of the image. Examples are shown in Figure 2. Here the smaller (regional) label feature encodes a pattern of ground pixels above water pixels, while the bigger (global) label feature encodes sky pixels at the top of the image, rhino/hippo pixels in the middle, and water pixels near the bottom. The global features can operate at a coarser resolution, specifying common value for a patch of sites in the label field. Our model learns these label features based on a set of labeled images.

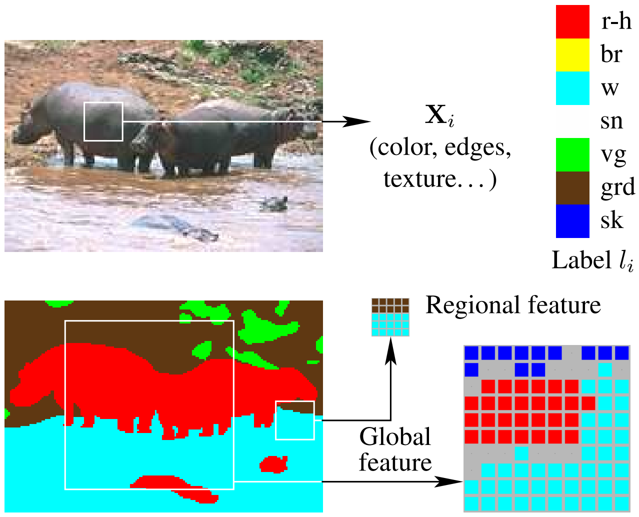

Figure 2: Above: An image patch at site i is represented by the outputs of several filters. The aim is to associate the patch with one of a predefined set of labels. Below: Example label features: regional (each cell corresponds to one site), which matches a boundary with ground (brown) above water (cyan); and global (each cell corresponds to 10×10 sites), which matches a rhino or hippo (red) in the water (cyan) with sky (blue) above the horizon. "Don't care" cells are blank (gray color). For label colors and abbreviations, see key in Fig. 5.

Associated with each label feature is a binary hidden variable, that acts as a switch for that feature. The feature encodes a particular label pattern through a parametrized conditional probability table (CPT) to the label sites within a region. This CPT specifies a multinomial probability distribution over the label values of each site. The hidden variables are assumed to be conditionally independent given the corresponding label variables, and vice versa (see Fig. 3). This structure has the form of a restricted Boltzmann machine (RBM) [5], in which inference and learning are greatly simplified.

Our multiscale conditional random field (mCRF) defines a conditional distribution over the label field $\mathbf{L} = \{l_i\}_{i \in S}$

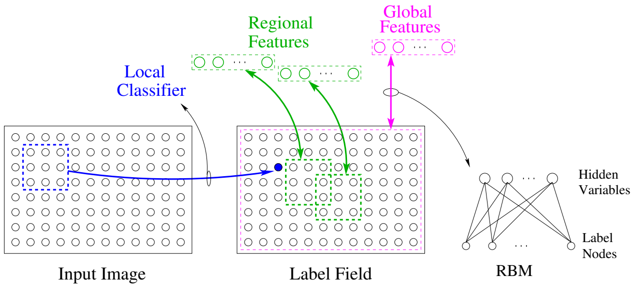

Figure 3: Graphical model representation. The local classifier maps image regions to label variables, while the hidden variables corresponding to regional and global features form an undirected model with the label variables. Note that features and labels are fully inter-connected, with no intra-layer connections (restricted Boltzmann machine).

given input image $\mathbf{X}$ by multiplicatively combining component conditional distributions that capture statistical structure at different spatial scales s:

$$P(\mathbf{L}|\mathbf{X}) = \frac{1}{Z}\prod_s P_s(\mathbf{L}|\mathbf{X}) \tag{1}$$

where $Z = \sum_L \prod_s P_s(\mathbf{L}|\mathbf{X})$ is a normalization factor (summed over all labelings). An mCRF is therefore a conditional form of the product-of-experts model [7]. Note that the model architecture makes the computation of Z tractable when conditioned on the image $\mathbf{X}$ and hidden variables, as the label field distribution can be factored across the sites given the values of the hidden variables and $\mathbf{X}$.

Each label feature in our model operates at a particular scale in the label field. For a given site, there are thus multiple predictors of its label conditioned on the image. Our model defined above, as in a standard CRF, combines the predictions of the various features multiplicatively. The product form of this combination has two chief effects on the system. First, label features need not specify the label of every site within the region. If a feature has uniform values for each possible label, it will play no role in the combination. We call this a "don't care" prediction. This enables a feature to focus its prediction on particular sites in the region. Second, the label of a site may be sharper than any of the component distributions. If two multinomials favor a particular value, then their product will be more sharply peaked on that value. Hence unconfident predictions that agree can produce a confident labeling.

In this paper, we instantiate the mCRF framework with three separate components, operating at three different scales s: a local classifier, regional features, and global features, as shown in Fig. 3.

1. Local Classifier. One powerful way of classifying a pixel of an image using information at a local level only is to use a statistical classifier, such as a neural network. Independently at each site i, the local classifier produces a distribution over label variable l_i given filter outputs $\mathbf{x}_i$ within an image patch centered on pixel i:

$$P_{\mathcal{C}}(\mathbf{L}|\mathbf{X},\boldsymbol{\lambda}) = \prod_i P_{\mathcal{C}}(l_i|\mathbf{x}_i,\boldsymbol{\lambda})$$

where $\boldsymbol{\lambda}$ are the classifier parameters. We use a multilayer perceptron as the classifier. Note that the classifier's performance is limited by class overlap and image noise [8].

2. Regional Label Features. This second component is intended to represent local geometric relationships between objects, such as edges, corners or T-junctions. Note that these are more than edge detectors: they specify the actual objects involved, thus avoiding impossible combinations such as a ground-above-sky border. We learn a collection of regional features from the training data.

We achieve a degree of translation invariance in the regional features by dividing the label field for the whole image into overlapping regions of the same size, on which these features are defined. The feature for a given region has its own hidden variables but share the CPT with other regions.

Let r index the regions, a index the different regional features within each region, and $j = \{1, \ldots, J\}$ index the label nodes (sites) within region r. The parameter $w_{a,j}$ connecting hidden regional variable $f_{r,a}$ and label node $l_{r,j}$ specifies preferences for the possible label value of $l_{r,j}$. So $w_{a,j}$ can be represented as a vector with $|\mathcal{L}|$ elements. We also represent the label variable $l_{r,j}$ as a vector with $|\mathcal{L}|$ elements, in which the vth element is 1 and the other is 0 when $l_{r,j} = v$. Thus, the probabilistic model describing regional label features has the following joint distribution:

$$P_{\mathcal{R}}(\mathbf{L},\mathbf{f}) \propto \exp\left\{\textstyle\sum_{r,a} f_{r,a}\mathbf{w}_a^T\mathbf{l}_r\right\}$$

where $\mathbf{f} = \{f_{r,a}\}$ represents all the binary hidden regional variables, $\mathbf{w}_a = [w_{a,1}, \ldots, w_{a,J}, \alpha_a]$, $\mathbf{l}_r = [l_{r,1}, \ldots, l_{r,J}, 1]$, and α_a is a bias term. Here the sites i are indexed by (r, j), because site i corresponds to a different node j in region r based on the position of that region in the image.

Intuitively, the most probable configuration of each feature is either the label pattern $\mathbf{l}_r$ in region r matching $\mathbf{w}_a$ and $f_{r,a} = 1$, or the label pattern $\mathbf{l}_r$ does not match $\mathbf{w}_a$ and $f_{r,a} = 0$. Given the hidden regional variables, the label variables are conditionally independent and the distribution of each label node can be written as

$$P_{\mathcal{R}}(l_i = v|\mathbf{f}) = \frac{\exp[\sum_{a,(r,j)=i} f_{r,a}w_{a,j,v}]}{\sum_{v'}\exp[\sum_{a,(r,j)=i} f_{r,a}w_{a,j,v'}]}$$

where the site is indexed by i and the summation ranges over all features defined on regions that contain i. Thus, the features specify a multinomial distribution over the label of each site. Finally, the regional component of our model is formed by marginalizing out the hidden variables in this sub-model: $P_{\mathcal{R}}(\mathbf{L}) \propto \prod_{r,a}[1+\exp(\mathbf{w}_a^T \mathbf{l}_r)]$.

3. Global Label Features. Each coarse-resolution *global feature* has as its domain the label field for the whole image (though in principle we could use smaller fields anchored at specific locations, as in Fig. 2). These global features are also configured as an RBM, with undirected links between the hidden global variables and the label variables. Let b index the global label patterns encoded in the parameters $\{\mathbf{u}_b\}$ and $\mathbf{g} = \{g_b\}$ be the binary hidden global variables. In order to encourage these variables to represent coarse aspects of the label field, we divide the label field into non-overlapping patches $p_m, m \in \{1, ..., M\}$, and for each hidden global variable g_b, its connections with the label nodes within patch p_m are assigned a single parameter vector u_{b,p_m}. These tied parameters effectively specify the same distribution for each label node within the patch (and reduce the number of free parameters). Like the regional component, the global label feature model has a joint distribution

$$P_{\mathcal{G}}(\mathbf{L}, \mathbf{g}) \propto \exp\left\{\sum_b g_b \mathbf{u}_b^T \mathbf{L}\right\}.$$

The global features also specify a multinomial distribution over each label node by their parameters. Note that a global feature, as well as a regional feature, can specify that it effectively "doesn't care" about the label of a given node or patch of nodes p, if its parameters $u_{bp}(v), v = 1, \ldots, |\mathcal{L}|$ are equal across label values v. This enables a feature to be sparse, and focus on labels in particular regions, allowing other features to determine the other labels. The joint model is marginalized to obtain the global feature component: $P_{\mathcal{G}}(\mathbf{L}) \propto \prod_b [1+\exp(\mathbf{u}_b^T \mathbf{L})]$.

4. Combining the Components. The multiplicatively combined probability distribution over the label field has a simple closed form (see Eqn. 1):

$$\begin{aligned} P(\mathbf{L}|\mathbf{X}; \boldsymbol{\theta}) = \tfrac{1}{Z} \textstyle\prod_i P_C^{\gamma}(l_i|\mathbf{x}_i, \boldsymbol{\lambda}) \times \\ \times \prod_{r,a}[1+\exp(\mathbf{w}_a^T \mathbf{l}_r)] \times \prod_b [1+\exp(\mathbf{u}_b^T \mathbf{L})] \end{aligned}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \{\mathbf{w}_a\}, \{\mathbf{u}_b\}, \gamma\}$ is the set of parameters in the model. We include a tradeoff parameter γ because the classifier is learned before the other components, and the model needs to modulate the effect of over-confident incorrect classifier outputs.

Equation 1 shows that the model forms redundant representations of the label field. A key attribute of our model, as in boosting and other expert combination approaches, is complementarity: each component should learn to focus on aspects modeled less well by others. Also, the labeling of an image must maximally satisfy all relevant predictions (the classifier's and the features') at every site. In particular, we expect the global features to help disambiguate (or even override) the classifier's judgment.

2.1. Parameter Estimation

For estimating the parameters $\boldsymbol{\theta}$, we assume a set of labeled images $D = \{(\mathbf{L}^t, \mathbf{X}^t), t = 1, ..., N\}$ is available. We train the conditional model discriminatively based on the Conditional Maximum Likelihood (CML) criterion, which maximizes the log conditional likelihood:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_t \log P(\mathbf{L}^t | \mathbf{X}^t; \boldsymbol{\theta}).$$

A gradient-based algorithm can be applied to maximize the conditional log likelihood. Calling h_s the unnormalized $P_s(\mathbf{L}|\mathbf{X})$, we obtain the following learning rule:

$$\Delta\boldsymbol{\theta}_s \propto \left\langle \frac{\partial \log h_s}{\partial \boldsymbol{\theta}_s} \right\rangle_{P_0(\mathbf{L}|\mathbf{X})} - \left\langle \frac{\partial \log h_s}{\partial \boldsymbol{\theta}_s} \right\rangle_{P_\theta(\mathbf{L}|\mathbf{X})}$$

where $\boldsymbol{\theta}_s$ are the parameters in component P_s, $P_0(\mathbf{L}|\mathbf{X})$ is the data distribution defined by D, and $P_{\boldsymbol{\theta}}(\mathbf{L}|\mathbf{X})$ is the model distribution. However, we need to calculate expectations under the model distribution, which is difficult due to the normalization factor Z. One possible approach is to approximate the expectations by Markov chain Monte Carlo (MCMC) sampling, but this requires extensive computation and the estimated gradients tend to be very noisy.

In this paper, we apply the contrastive divergence (CD) algorithm [7]. CD is an approximate learning method that overcomes the difficulty of computing expectations under the model distribution. The key benefit of applying CD to learning parameters in a random field is that rather than requiring convergence to equilibrium, such as in MCMC, one only needs to take a few steps in the Markov chain to approximate the gradients, which can be a huge savings, particularly during learning when the gradients must be updated repeatedly. In addition, because our model is a form of additive random field, a block Gibbs sampling chain can be implemented efficiently, simply computing the conditional probabilities of the feature sets $\mathbf{f}$ and $\mathbf{g}$ given $\mathbf{L}$ and vice versa. The original CD algorithm optimizes the parameters of a model by approximately maximizing data likelihood; we extend it here to the objective of maximizing conditional likelihood.

2.2. Inference for Labeling an Image

To label a new image $\mathbf{X}$, we need to infer the optimal label configuration $\mathbf{L}$ given $\mathbf{X}$. There are two main criteria for inferring labels from the posterior distribution [1]: maximum a posteriori (MAP) and maximum posterior marginals (MPM). Exact MAP is difficult to compute due to the high dimensionality and discrete domain of $\mathbf{L}$. Also, it can be too conservative in searching approximate solutions because it only considers the most probable case and disregards the difference between other solutions. The MPM criterion, which minimizes the expected number of the mislabeled sites by taking the modes of posterior marginals:

$$l_i^* = \arg\max_{l_i} P(l_i|\mathbf{X}), \qquad \forall i \in S$$

usually produces a better solution. In this paper, we adopt MPM. Evaluating $P(l_i|\mathbf{X})$ in our model is intractable due to its loopy structure, so we must resort to approximate inference methods. We use Gibbs sampling due to its simplicity and fast convergence. Note that we can take advantage of our architecture to start sampling the chain in a reasonable initial point, given by the label distribution output by the classifier.

3. Experimental Results

3.1. Data Sets

We applied our mCRF to two natural image datasets. The first dataset is a 100-image subset of the Corel image database, consisting of African and Arctic wildlife natural scenes. We labeled them manually into 7 classes: 'rhino/hippo', 'polar bear', 'vegetation', 'sky', 'water', 'snow' and 'ground'. The training set includes 60 randomly selected images and the remaining 40 for testing; each image is 180×120 pixels.

The second dataset, the Sowerby Image Database of British Aerospace, is a set of color images of out-door scenes and their associated labels. The images contain many typical objects near roads in rural and suburban area. After preprocessing the images as in [3], we obtain 104 images with 8 labels: 'sky', 'vegetation', 'road marking', 'road surface', 'building', 'street objects', 'cars' and 'unlabeled'. During testing, we do not consider the unlabeled sites and the model's output for them. We randomly select 60 images as training data and use the remaining 44 for testing; each image is 96×64 pixels.

We extract a set of image statistics $\mathbf{x}_i$ at each image site i, including color, edge and texture information. In these experiments, each site corresponds to a single image pixel. For the color information, we transform the RGB values into CIE Lab* color space, which is perceptually uniform. The edge and texture are extracted by a set of filterbanks including difference-of-Gaussian filter at 3 different scales, and quadrature pairs of oriented even- and odd-symmetric filters at 4 orientations $(0, \pi/4, \pi/2, 3\pi/4)$ and 3 scales. Thus each pixel is represented by 30 image statistics.

3.2. Model Training

We train the system sequentially: first we train the local classifier; then we fix the classifier and train the label features. Although potentially suboptimal with respect to a joint training of all parameters, the sequential approach is more efficient. The classifier is a 3-layer multilayer perceptron (MLP) with sigmoid hidden units and $|\mathcal{L}|$ outputs with softmax activation function (so we can interpret the output as the posterior distribution over labels). For each image site, the input of the MLP is the image statistics within a local 3×3 pixel window centered at that site. Larger window sizes (e.g. 5×5) produced only small improvements in the classification rate but need much longer training. The MLP is trained to minimize the cross-entropy for multiple classes with a scaled conjugate gradient algorithm. In the CD algorithm, we always run a Markov chain for 3 steps from the correct label configuration.

We compare our approach with an MRF, defined as a generative model $P(\mathbf{L}, \mathbf{X}) = \prod_i P(\mathbf{x}_i|l_i)P(\mathbf{L})$, where $\mathbf{x}_i$ is the image statistics vector at image site i. The class-conditional density $P(\mathbf{x}_i|l_i)$ is modeled by a Gaussian mixture with diagonal covariance matrices. We learn the Gaussian mixture with the EM algorithm and choose the number of mixture components using a validation set. The label field $P(\mathbf{L})$ is modeled by a homogeneous random field defined on a lattice:

$$P(\mathbf{L}) \propto \exp\left\{ -\sum_{i \in S} \sum_{j \in \mathcal{N}_i} \sum_{u,v} \mu_{u,v} \delta(l_i - u)\delta(l_j - v) \right\}$$

where the parameter $\mu_{u,v}$ measures the compatibility between neighboring nodes (l_i, l_j) when they take the value (u, v). We trained the random field model $P(\mathbf{L})$ using the pseudo-likelihood algorithm [12]. To infer the optimal labeling given a new image, we use the same MPM criterion where the marginal distribution is calculated by the loopy belief propagation algorithm [4].

3.3. Performance Evaluation

We evaluate the performance of our model by comparing with the generative MRF and the local classifier over the Sowerby and Corel datasets. The correct classification rates on the test sets of both datasets are shown in Table 1. For the Corel dataset, the local classifier is an MLP with 80 hidden nodes, and the regional features are defined on 8×8 regions with overlap 4 in each direction, while the global features are defined on the whole label field with patch size 18×12. There are 30 regional features and 15 global features. For the Sowerby data, the local classifier has 50 hidden units and the regional features are defined on 6×4 regions overlapped by 2 horizontally and 3 vertically. The global features are defined on 8×8 patches of label sites. There are 10 global features and 20 regional features. For both mCRFs, we set the classifier weighting parameter ($\gamma = 0.9$) and the model structure—number of regional and global features, and region sizes—using a small validation set.

Table 1: Classification rates for the models.

Database	Classifier	MRF	mCRF
Corel	66.9%	66.2%	80.0%
Sowerby	82.4%	81.8%	89.5%

Table 2: Confusion matrix in percentage for Corel data. Entry (row i, column j) means true label i was estimated as j.

	r-h ■	br ■	w ■	sn	vg ■	grd ■	sk ■
■	9.27	0.14	0.53	0.01	1.01	1.00	0
■	0.08	8.06	0.01	0.52	0.12	0.63	0
■	0.33	0	12.87	0	0.42	0.76	0.05
	0	0.82	0	12.83	0.23	0.09	0.04
■	0.95	0.55	0.09	3.18	15.06	2.99	0.06
■	1.13	1.18	1.11	0.26	1.56	21.19	0
■	0	0	0	0	0.19	0.01	0.66

From Table 1, we can see that the performance of the MLP classifier is comparable to the MRF, while our model provides a significant improvement. The result shows the advantage of discriminative over generative modeling and the weakness of local interactions captured by the MRF model. The confusion matrix for the testing results on our mCRF model is shown in Tables 2–3, where the values show the percentage of labels in the whole testing data. The tables show that the errors made by our model are consistent across the classes. For the Sowerby data, the overall performance is comparable to the best result in published classification result on this dataset: 90.7% in [13].

We also show the outputs of the local classifier, MRF and our model on some test images in Figure 5. The classifier works reasonably well but can be easily fooled since no contextual information is included. The MRF produces quite smooth label configurations but it may smooth in a wrong way because it captures only local context, which can be misleading. Our mCRF model generates more reasonable labelings in which the contextual information provided by regional and global features corrects most of the wrong predictions from the local classifier—even when these occupy large, scattered portions in the image. We can take the probability of labeling for each site as a confidence measure, and form a confidence map of the labeling (see Fig. 5, rightmost column). This confidence measures the quality of the prediction in a consistent way: note how it tends to be low around boundaries and where the model cannot reverse the classifier's wrong labeling due to confusion by highlights or shadows. The model performance in this case could be improved by letting the label features have access to image statistics as well.

Table 3: Confusion matrix in percentage for Sowerby data.

	sk ■	vg ■	rdm	rds ■	bd ■	str ■	car ■
■	12.01	0.53	0.00	0.01	0.03	0.00	0.01
■	0.83	33.39	0.01	1.41	2.71	0.03	0.09
	0.00	0.00	0.08	0.10	0.00	0	0
■	0.01	0.94	0.02	40.33	0.10	0.01	0.05
■	0.06	2.60	0.02	0.30	3.05	0.01	0.05
■	0.02	0.25	0	0.03	0.12	0.02	0.01
■	0.02	0.27	0.00	0.09	0.24	0.00	0.14

Figure 4 shows a subset of the parameters learned, i.e., the conditional probability tables in the regional and global features. For legibility, only the most probable labels are shown for each site and each feature pattern is displayed as a matrix of blocks. The color of each block represents the label value with the highest probability (cf. the key in Fig. 5) and the block size is proportional to the probability values. Figure 4 shows 5 regional features from the Sowerby data and 5 global features from the Corel data. We can see that the regional features capture within-label regularities as well as cross-label boundary regularities. For example, the first regional feature is mostly devoted to 'ground', while the fourth one represents the boundary between 'vegetation' and 'sky'. The global features capture coarser patterns in the entire label field and reflect the global context in the data. For instance, the second global feature shows the rhino or hippo is usually surrounded by vegetation and water, and the sky is above them, while the fourth one shows the bear is often surrounded by snow.

4. Discussion

The method proposed here is similar to earlier approaches to the problem of object detection, or the more general task of image labeling, in that it combines local classifiers with probabilistic models of label relationships. Insight into these various models can be gained by comparing the solutions to the basic problem posed in the introduction: how can information at different scales be represented, learned,

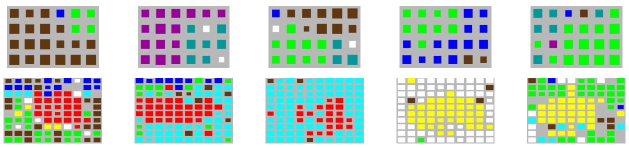

Figure 4: Examples of learned regional label features from the Sowerby dataset (above, 6×4 sites) and global label features on the Corel dataset (below, 10×10 blocks each of 18×12 sites).

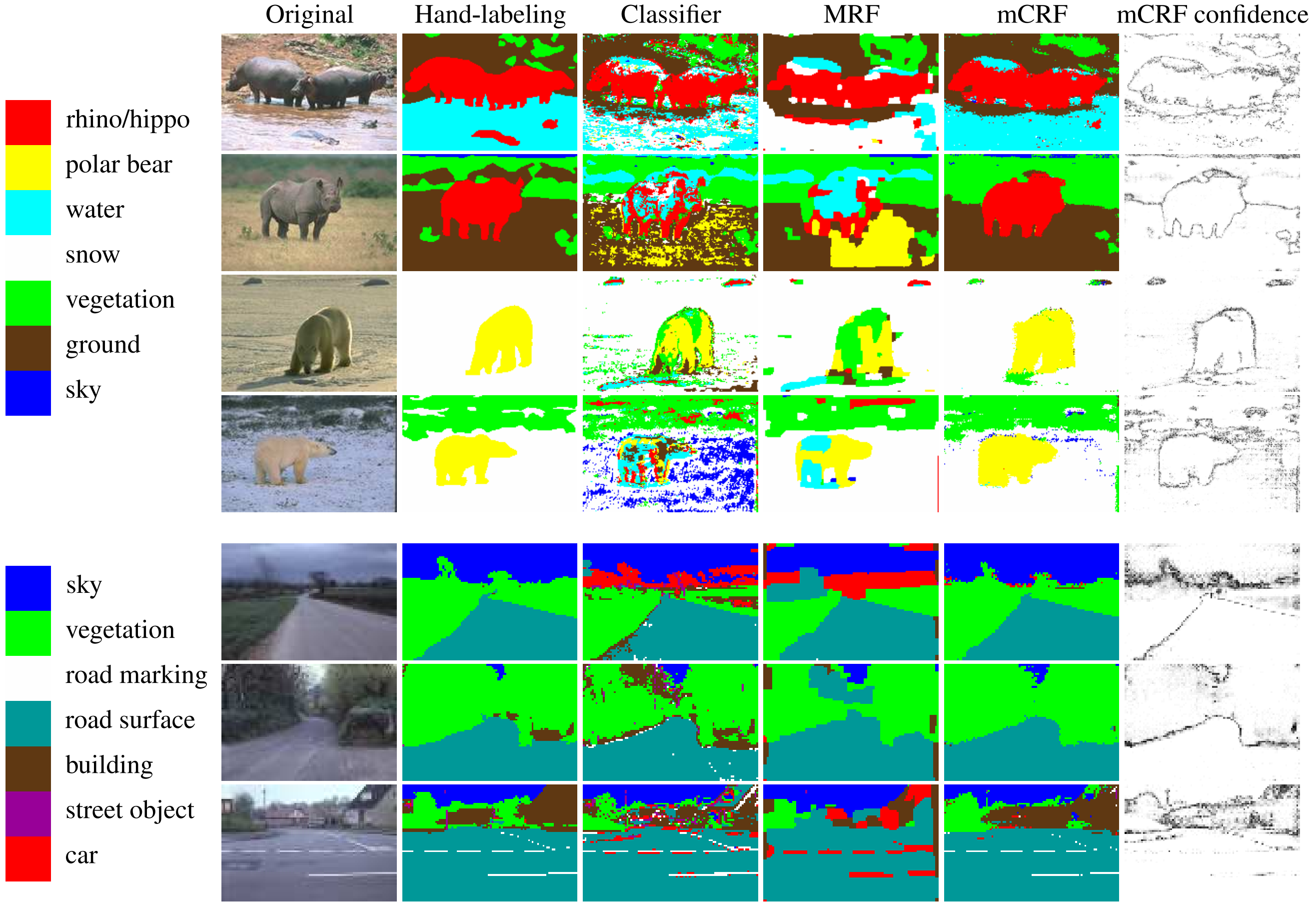

Figure 5: Some labeling results for the Corel (4 top rows) and Sowerby (3 bottom rows) datasets, using the classifier, MRF and mCRF models. The color keys for the labels are on the left. The mCRF confidence is low/high in the dark/bright areas.

and combined?

A primary difference between these earlier models and our model is the form of the representation over labels. One method of capturing label relationships is through a more conceptual graphical model, such as an abstraction hierarchy consisting of scenes, objects, and features [14]. The distribution over labels can also be obtained based on *pairwise* relationships between labels at different sites. Recently, Kumar and Hebert [9] extended earlier MRF approaches [6] by including image information in the learned pairwise compatibilities between labels of different sites. Training their model discriminatively as opposed to generatively led to significant improvements in detecting man-made structures in images over traditional MRF approaches.

An alternative to a pairwise label model is a tree-based model [3, 13]. Tree-based models have the potential to represent label relationships at different scales, corresponding to conditional probability tables at different levels in the tree. Static tree-based models are limited in their flexibility due to the fixed nature of the tree, which tends to lead to blocky labelings. The dynamic tree model [13] elegantly overcomes this approach by constructing the tree on-line for a given image; however, inference is quite complicated in this model, necessitating complicated variational techniques. Thus the CPTs learned in this model were restricted to very simple label relationships.

In our model, a wide variety of patterns of labels, at different scales, are represented by the features, and the features all interact at the label layer. The mCRF model is flatter than the trees, and the features redundantly specify label predictions. The model is therefore searching for a single labeling for a given image that maximally satisfies the constraints imposed by the active learned features. In the tree-based models, alternative hypotheses are represented as different trees, and inference considers distributions over trees. Our method instead combines the probabilistic predictions of different features at various scales using a product model, which naturally takes into account the confidence of each feature's prediction.

Our model is an instantiation of a larger framework, where individual sub-models specialize on tasks and have access to particular information. Further work can consider, for example, label features over a range of scales (rather than just local and global), or label features that have also access to some image statistics. Generative models cannot include image information as well as label patterns into learned features. We expect that this will enable the features to localize boundaries between objects in a more precise manner.

Ideally the system we described would be applied to a higher level of input image representation, to apply to labeled image features rather than individual pixels. However, this requires a consistent and reliable method for extracting such representations from images.

Finally, automatic image labeling has several direct applications, including video surveying or object detection and tracking. A primary application is content-based image retrieval. Many current content-based query methods rely on global image properties, which do not handle searches for specific objects in a variety of scenes [2]. As the quality of image data increases, it becomes more important to have a mechanism for classifying images as fully as possible prior to insertion into a database. After learning our model on a small, representative data set, the entire database can be labeled automatically. Then, user queries such as "find images with hippos in water" can be processed very quickly. Indexes for the classes associated with each image could be generated for each image, which would allow rapid retrieval; alternatively, more specific regions of images can be retrieved based on the pixel labels.

5. Conclusions

We have presented a novel probabilistic model for labeling images into a predefined set of class labels. The model is a product combination of individual models, each providing labeling information from different aspects of the image: a classifier that looks at local image statistics; regional label features that look at local label patterns; and global label features that look at large, coarse label patterns. Both the classifier and the label features are learned from a training set of labeled images. This strategy results in consensual labelings that have to agree with the image statistics but at the same time respect geometric relationships between objects at a local and global scale. The main reasons for our model's success are its direct representation of large-scale interactions and its devoting resources to modelling the label space but not the image space. A chief novelty of the work is that we generalize the standard form of feature functions used in CRFs to use hidden variables, each encoding a learned pattern within a subset of label variables.

Acknowledgments

We thank Max Welling and Geoff Hinton for discussions on contrastive divergence, the anonymous reviewers, and BAE Systems for letting us use their Sowerby Image Database. Funded by grants from CIHR New Emerging Teams program and the Institute for Robotics and Intelligent Systems.

References

[1] C. Bouman and M. Shapiro: "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Processing* 3:162–177, 1994.

[2] N. W. Campbell, W. P. J. Mackeown, B. T. Thomas, and T. Troscianko: "Interpreting image databases by region classification," *Pattern Recognition* 30:555–563, 1997.

[3] X. Feng, C. K. I. Williams, and S. Felderhof: "Combining belief networks and neural networks for scene segmentation," *IEEE Trans. PAMI* 24:467–483, 2002.

[4] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael: "Learning low-level vision," *Int. J. Comp. Vision* 40:25–47, 2000.

[5] Y. Freund and D. Haussler: "Unsupervised learning of distributions on binary vectors using 2–layer networks," *NIPS*, 1992.

[6] S. Geman and D. Geman: "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. PAMI* 6:721–741, 1984.

[7] G. E. Hinton: "Training products of experts by minimizing contrastive divergence," *Neural Comp.* 14:1771–1800, 2002.

[8] S. Konishi and A. L. Yuille: "Statistical cues for domain specific image segmentation with performance analysis," *CVPR*, pp. 125–132, 2000.

[9] S. Kumar and M. Hebert: "Discriminative random fields: A discriminative framework for contextual interaction in classification," *ICCV*, pp. 1150–1157, 2003.

[10] J.-M. Laferte, F. Heitz, P. Perez, and E. Fabre: "Hierarchical statistical models for the fusion of multiresolution data," *ICCV*, 1995.

[11] J. Lafferty, A. McCallum, and F. Pereira: "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *ICML*, pp. 282–289, 2001.

[12] S. Z. Li: *Markov Random Field Modeling in Image Analysis*. Springer, 2001.

[13] A. J. Storkey and C. K. I. Williams: "Image modelling with position encoding dynamic trees," *IEEE Trans. PAMI* 25:859–871, 2003.

[14] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin: "Context-based vision system for place and object recognition," *ICCV*, pp. 273–280, 2003.