An Empirical Bayes Approach to Contextual Region Classification

Svetlana Lazebnik
University of North Carolina
Chapel Hill, NC 27599
lazebnik@cs.unc.edu

Maxim Raginsky
Duke University
Durham, NC 27708
m.raginsky@duke.edu

Abstract

This paper presents a nonparametric approach to labeling of local image regions that is inspired by recent developments in information-theoretic denoising. The chief novelty of this approach rests in its ability to derive an unsupervised contextual prior over image classes from unlabeled test data. Labeled training data is needed only to learn a local appearance model for image patches (although additional supervisory information can optionally be incorporated when it is available). Instead of assuming a parametric prior such as a Markov random field for the class labels, the proposed approach uses the empirical Bayes technique of statistical inversion *to recover a contextual model directly from the test data, either as a spatially varying or as a globally constant prior distribution over the classes in the image. Results on two challenging datasets convincingly demonstrate that useful contextual information can indeed be learned from unlabeled data.*

1. Introduction

This paper considers the problem of local region labeling in images: we want to classify every pixel (or small patch) by its semantic class, e.g., sky, grass, water, building, etc. This task is challenging because of the well-known "aperture problem" of local ambiguity: for example, a uniformly blue image patch may be a piece of sky, calm water, or a painted wall. To resolve this ambiguity, it is necessary to look at the patch within the context of a larger image area surrounding it. Therefore, a good approach to local region classification must incorporate a contextual model that captures the probability of different classes occurring nearby, or sharing a specific spatial relationship. Recent literature contains many approaches for contextual image labeling [4, 6, 11, 18, 19, 20, 21]. Most existing contextual models must be learned from training data that contains a representative sampling of all possible inter-class interactions. Since the number of even pairwise interactions is quadratic in the number of classes, it is usually difficult to get enough data to estimate an expressive context model. One traditional way to deal with this difficulty is to adopt a simple parametric prior that only encodes generic assumptions about the smoothness of the label field — e.g., a Markov Random Field with Potts potentials [8]. Recently, some researchers have suggested more creative ways of dealing with sparse training data, such as mining semantic knowledge about class associations from the Web [11].

In this paper, instead of making strong smoothness assumptions or searching for external sources of knowledge to help with context estimation, we ask whether there is any low-level information *intrinsic to the unlabeled test images* that can allow us to learn a contextual model over class labels. We are specifically interested in the scenario where the labeled training data may be sufficient for learning what small patches from each class look like locally, but not for observing all the possible ways in which patches from different classes can co-occur within larger neighborhoods. In this situation, deriving a contextual model directly from the test data would be very useful because, after all, the test data contains precisely the inter-class interactions that we have to get right! Our work can also be viewed as a principled attempt to address the problem first posed by Divvala et al. [3]: starting with an imperfect region classification model learned from limited training data, can we improve its performance by leveraging the wealth of unlabeled data in the potentially much larger test set?

How can it be possible to deduce priors over class label sequences from an unlabeled image collection? At first glance, it may seem counter-intuitive that this can be done at all. However, provided there is a sufficiently stable stochastic relationship between class labels of individual local regions and corresponding image observations, then the sequence of image observations forms an indirect and noisy reflection of the unseen class label sequence and retains a lot of information about the structure of that sequence. In effect, if we know the process that converts local labels to image features, and if we observe a sequence of image features, we should be able to go back from the features to the labels. This insight can be formally captured with the help of *empirical Bayes* methodology from statistics literature [13, 14], in which priors are inferred from data instead

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

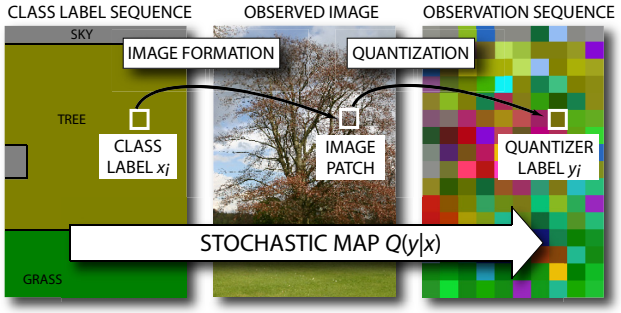

Figure 1. Image formation model used in this work. Image observations (in our case, quantized image features such as texture and color) are produced by a stochastic or "noisy" mapping Q applied to the underlying true image labels. We view region classification as the process of *denoising* the observation sequence to recover the underlying clean label sequence.

of being specified in advance. This methodology has recently given rise to an information-theoretic framework of *universal denoising* [22], which, in turn, has inspired our own work. We think of the stochastic mapping from class labels to observations as a "noisy channel," and then we infer the underlying class labels sequence by denoising the observation sequence (Figure 1).

Figure 2 gives a pictorial overview of the approach proposed in this paper. We use the training data only to learn the *local likelihood model*, or the conditional probabilities of patch-level observations given class labels. Intuitively, the local likelihood model captures the appearances of individual classes, which are the basic "building blocks" of scenes. The knowledge of the local likelihood model then allows us to recover a contextual prior over a sequence of test observations using the empirical Bayes technique of *statistical inversion*. Finally, the contextual prior is combined with the local likelihood to perform Maximum a Posteriori (MAP) region classification. This approach will be formally described in the text section, followed in Section 3 by implementation details and experimental results.

2. The Approach

Figure 1 illustrates the basic "generative" framework followed in this paper. Underlying each image is a sequence or a field of class labels. Crucially, we do not assume any parametric prior model (such as an MRF) for this sequence. Instead, our priors will be deduced in a data-driven fashion by looking at the statistical regularities in the observations. Each class label x generates an image patch through some unmodeled image formation process, and each patch in turn undergoes a feature extraction and quantization step to produce a discrete observation y (the particular features and quantization procedures used in our work will be described in Section 3.1). The composition of the image formation and feature extraction steps gives rise to a stochastic mapping from class labels x to observations y. This mapping is fully described by the likelihood model of observations given underlying classes. We can represent these likeli-

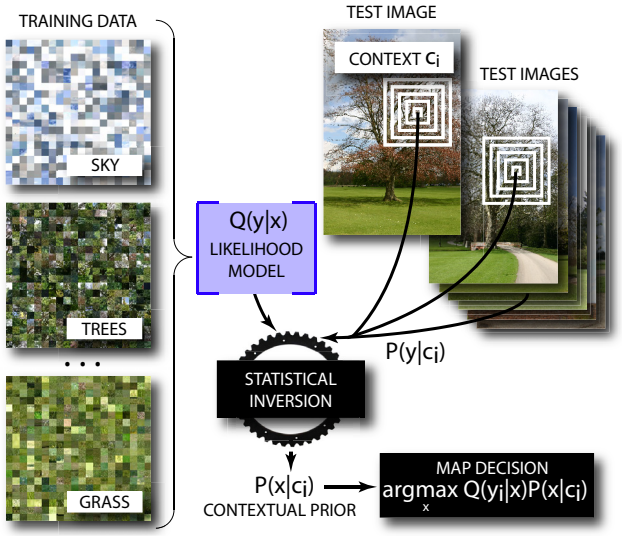

Figure 2. Overview of the proposed approach. At training time, all we need to learn is a likelihood model $Q(y|x)$ for local image patches y given class labels x – that is, we only need to know what image patches from different scene classes look like locally, not how they co-occur. The contextual information is recovered directly from the unlabeled test images by exploiting the statistical redundancy of the observations. We group observations that occur in similar neighborhood contexts and obtain an estimate of the distribution of underlying class labels in each group by *statistical inversion* (Section 2). This yields a contextual prior that is combined with the local likelihood to perform MAP classification of regions.

hoods in the form of a matrix Q each of whose rows $Q(\cdot|x)$ is the conditional probability distribution of the observation Y given x. The training stage of our approach consists of learning the likelihood model from a set of training observations paired with their class labels.

Next, we need to know how to perform region classification on a set of previously unseen test images. These test images correspond to a sequence of observations $\mathbf{y}$, and our job is to infer the underlying class label sequence $\mathbf{x}$. We conceptualize this problem as *denoising*, or recovery of a random sequence whose elements are independently corrupted by a known *noisy channel* (i.e., the stochastic mapping Q), where we seek to minimize the expected fraction of incorrectly recovered symbols. Denoising can be formulated as a *compound decision problem* [12], or a set of simultaneous statistical decision problems that have some shared structure. In our case, each separate problem is the recovery of a correct class label for a single test observation, and the shared structure comes from the assumption that the local stochastic mapping from the "clean" sequence of class labels to the "noisy" sequence of observations is the same for all image locations.

The goal of constructing robust and tractable decision procedures that perform well without being tuned to a specific parametric model has given rise to a powerful class of nonparametric *universal denoising* techniques [22], which can asymptotically approach optimal performance among all sliding-window schemes on any stationary ergodic ran-

dom field. The idea behind universal denoising is to use empirical frequencies of observed noisy symbols conditioned on contexts computed from other observations in their neighborhood to infer the corresponding frequencies of the underlying clean symbols.

In the following, we adapt the methodology of [22] to derive a denoising-type procedure for region classification. Formally, a *compound decision procedure* is a rule for obtaining the estimates $\hat{x}_i$ of the clean symbols x_i, $i = 1, \ldots, n$, based on the *entire* observation sequence $\mathbf{y}$, i.e., $\hat{x}_i = f_i(\mathbf{y})$ for some function f_i. The goal is to make the average probability of error $\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left[f_i(\mathbf{Y}) \neq X_i\right]$ as small as possible without assuming a specific form for the prior distribution for $\mathbf{x}$.

According to Bayesian decision theory [14], the optimal estimate of x_i is the one that has maximal posterior probability given the *entire* test sequence:

$$\hat{x}_i = \arg\max_x \mathbb{P}(X_i = x | \mathbf{Y} = \mathbf{y}) .$$

Unfortunately, estimating label posteriors conditioned on all possible test sequences is not feasible. In order to arrive at a more tractable and realistic decision procedure, we need to make some simplifications. To begin with, we will restrict the compound decision procedures under consideration to sliding-window rules that "see" observations in a suitably defined neighborhood of a given patch. This restriction is realistic, provided the neighborhood is large enough to capture most of the contextual influences on x_i. For each i, let $\mathbf{y}_{\mathcal{N}(i)}$ denote the vector of observations in i's neighborhood, let $\mathbf{y}_i = (y_i, \mathbf{y}_{\mathcal{N}(i)})$, and suppose that an estimate of x_i is given by $\hat{x}_i = f_i(\mathbf{y}_i)$. Now the problem of optimal estimation of x_i hinges on the estimation of the posterior $\mathbb{P}(X_i = x_i | \mathbf{Y}_i = \mathbf{y}_i)$. We can further assume that each observation y_i is conditionally independent of all the other observations given its label x_i (which corresponds to the factorized likelihood model adopted by many MRF approaches to image labeling [8]). With this assumption, we can write down a simplified posterior probability:

$$\begin{aligned} \mathbb{P}(x_i|\mathbf{y}_i) &\propto \mathbb{P}(y_i|x_i, \mathbf{y}_{\mathcal{N}(i)})\, \mathbb{P}(x_i|\mathbf{y}_{\mathcal{N}(i)}) \\ &= Q(y_i|x_i)\, \mathbb{P}(x_i|\mathbf{y}_{\mathcal{N}(i)}) . \end{aligned}$$

Because there is still a very large number of possible neighborhood sequences $\mathbf{y}_{\mathcal{N}(i)}$, we make our model more flexible by conditioning on *contexts* c_i, or functions of $\mathbf{y}_{\mathcal{N}(i)}$ defined to make the estimation task easier. For example, a very simple kind of context function would be a histogram of symbols in $\mathbf{y}_{\mathcal{N}(i)}$. In Section 3, we will discuss the actual context functions used in this work. Once a suitable context function has been defined, we can replace $\mathbb{P}(x_i|\mathbf{y}_{\mathcal{N}(i)})$ by $\mathbb{P}(x_i|c_i)$ and define our decision procedure by

$$\hat{x}_i = \arg\max_x Q(y_i|x)\hat{P}(x|c_i) , \qquad (1)$$

Supervised (training) phase: learn local likelihood model $Q(y|x)$ on a training set of labeled patches.

Unsupervised (test) phase:

1. Extract observation sequence $\mathbf{y}$ from the test images.
2. For each observation y_i, compute context c_i.
3. For each test patch i:
   - Estimate empirical distribution $\hat{P}(y|c_i)$ from the entire test sequence.
   - Obtain *contextual prior* $\hat{P}(x|c_i)$ by statistical inversion (eq. 2).
   - Find $\hat{x}_i$ by MAP rule: $\hat{x}_i = \arg\max_x Q(y_i|x)\hat{P}(x|c_i)$.

Table 1. Summary of the empirical Bayes approach to contextual region classification (see also Figure 2).

where $\hat{P}(x|c_i)$ is an empirical estimate of $\mathbb{P}(x_i|c_i)$. But how can we obtain this estimate if we do not actually observe the values of x at test time? We can find the empirical distribution $\hat{P}(y|c_i)$ of the observations given their contexts, but how can we go from $\hat{P}(y|c_i)$ to $\hat{P}(x|c_i)$ without any strong assumptions on the distribution of x?

This is where we need to bring in the idea of *statistical inversion*. We have

$$\mathbb{P}(Y = y|c) = \sum_x Q(y|x)\, \mathbb{P}(X = x|c) .$$

Let $\hat{P}_{X|c}$ and $\hat{P}_{Y|c}$ be column vectors representing empirical estimates of $\mathbb{P}(X|c)$ and $\mathbb{P}(Y|c)$, respectively – and at this point, we can only obtain $\hat{P}_{Y|c}$, while $\hat{P}_{X|c}$ remains an unknown quantity. By the law of large numbers, we can write $\hat{P}_{Y|c} \approx Q^T \hat{P}_{X|c}$. In principle, this is a linear equation that can be solved for $\hat{P}_{X|c}$. It can also be shown that $Q^{-T}\hat{P}_{Y|c}$ (where Q^{-T} is the Moore-Penrose pseudoinverse of Q^T) is a *statistically consistent* estimate of $\hat{P}_{X|c}$, i.e., an estimate that converges to the true distribution with probability one [22]. However, literal inversion is not guaranteed to be consistent for small sample sizes, possibly resulting in solutions that have negative components and do not sum to one. Instead, we obtain $\hat{P}_{X|c}$ by minimizing over all clean distributions P_X the Kullback-Leibler divergence between $\hat{P}_{Y|c}$, the empirical noisy distribution, and $Q^T P_X$, the distribution obtained by applying the stochastic map given by Q to the clean distribution P_X:

$$\hat{P}_{X|c} = \arg\min_{P_X} D(\hat{P}_{Y|c} \| Q^T P_X) . \qquad (2)$$

This optimization problem is convex, and can be solved iteratively by Expectation-Maximization (EM). For completeness, the update equation is as follows:

$$P_{X|c}^{(t+1)}(x) = P_{X|c}^{(t)}(x) \sum_y \frac{Q(y|x)\hat{P}_{Y|c}(y)}{\sum_{x'} P_{X|c}^{(t)}(x') Q(y|x')} .$$

Now all the ingredients are in place, and we can write down a very simple procedure for converting the output sequence $\mathbf{y}$ into an estimated clean sequence. This procedure, given

in Table 1, implements the MAP rule (1) where the likelihoods $Q(y_i|x)$ are learned from the training data, while the contextual priors $\hat{P}(x|c_i)$ are estimated from the test sequence via statistical inversion.

3. Implementation and experiments

This section describes our implementation of the proposed empirical Bayes scheme. Our main evaluation platform is the standard MSRC dataset [19], which includes 21 common scene classes such as sky, grass, buildings, people, dogs, etc. In Section 3.6, we will show additional results on the geometric context dataset of Hoiem et al. [6].

3.1. Local likelihood model

We perform feature extraction by dividing the images into non-overlapping 20×20 pixel patches and computing four types of features from each patch: position, SIFT [9], textons, and color. Textons are computed by convolving the images with a filter bank and recording the index of the filter with the maximum absolute response at each pixel. The texton descriptor is the histogram of texton indices within the patch. We use a subset of the LM filter bank [7] consisting of 18 second-derivative-of-Gaussian and 6 Laplacian filters, and 13 filters from the S filter bank [17], for a total of 37 filters. Because we distinguish between positive and negative filter responses, the texton histogram has 74 dimensions. For color, we compute a 48-dimensional descriptor by subdividing the patch into a 4×4 grid and finding the mean color (in the CIE Lab space) of each grid cell. Our feature extraction scheme is similar to that of Verbeek and Triggs [21], except that they have a more sophisticated color descriptor and no texton channel.

We quantize SIFT, texton, and color descriptors using visual codebooks of 500 centers each learned from training data using k-means. Position is quantized using a uniform 5×5 grid superimposed over the image. To cope with the $500^3 \times 25$ possible distinct output index combinations, we make the usual Naive Bayes assumption that the feature channels are conditionally independent given the underlying class label x. Then the likelihood model is given by $Q(y|x) = \prod_m Q^m(y^m|x)$, where $Q^m(y^m|x)$ is the stochastic mapping between x and the mth feature type, and each Q^m is estimated separately by smoothed empirical counts from the labeled training set. Table 2 reports maximum likelihood classification results on the MSRC dataset using this model with different combinations of cues (here and in all the following, we use the standard split of this dataset into 276 training and 256 test images [19]). The table confirms that using all four cues together is beneficial.

It must be noted that the above vocabulary-based feature extraction scheme is just one of many possible schemes that can be used with the general framework of Table 1. Other discrete observation models, such as randomized forests [18, 20], can be accommodated just as easily.

Feature combination	Per-patch	Per-class
SIFT only	24.04	18.63
SIFT + Texton	32.42	24.86
SIFT + Texton + Color	50.02	37.31
SIFT + Texton + Color + Pos	**53.26**	**40.65**

Table 2. The performance of the local likelihood model on the MSRC dataset. Following standard practice, we report both the overall per-patch classification rate (i.e., the percentage of all regions correctly classified) and the average of per-class classification rates.

3.2. Neighborhood contexts

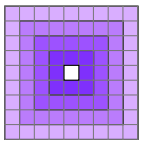

Figure 3. Concentric context.

We now discuss our representation of the context function. We have experimented with several ways of subdividing a square neighborhood around a central patch, looking for the right tradeoff between spatial selectivity and invariance to permutations of context elements. The best performance was obtained by subdividing the neighborhood into concentric "rings" (Figure 3). The context function for a given neighborhood is computed by taking the marginal histograms of SIFT, color, and texton output labels within each "ring" and concatenating them together, resulting in a high-dimensional but sparse vector of counts. Note that since the histograms within each "ring" are normalized, the observations that occur closer to the center are weighted more heavily in the context.

Having defined the form of the context function, we next need to be able to estimate $\hat{P}(y|c)$, the probability that a given observation y occurs in the center of a neighborhood with context c. This is a challenging task, as it is very rare to observe multiple identical contexts in the image sequence. To deal with this, we use a k-nearest neighbor estimate: For each region i, we find $k = 500$ other regions whose contexts are the most similar to c_i, where the similarity function is given by histogram intersection which can be computed very quickly for sparse vectors, and let $\hat{P}(y|c_i)$ be the empirical distribution of observations y over these contexts. Currently, we do exhaustive nearest neighbor search (for the MSRC dataset, this amounts to classification time of about three to five seconds per image, depending on the dimensionality of the context), but in the future we plan to explore fast approximate search [16] to speed up this task.

Table 3(a) shows the performance of our classification scheme using the neighborhood context representation. Compared to the local likelihood results shown in Table 2, incorporating the contextual prior improves the overall accuracy by over 10%. This confirms that the unlabeled test set does indeed contain useful contextual information that can be used to improve performance over purely local region classification. As a basic check, we would also like to know how unsupervised contextual priors compare to supervised ones, when the training data allows us to estimate them. Table 3(b) shows the performance achieved on the

Context type	Per-patch	Per-class
(a) Neighborhood (unsupervised)	63.65	52.68
(b) Neighborhood (supervised)	64.60	52.63
(c) Image-level (unsupervised)	66.20	56.49
(d) Neighborhood + image (unsupervised)	69.14	59.63
(e) Unsupervised + supervised	**72.14**	**62.80**

Table 3. Contextual classification performance for the MSRC dataset using different context models (see text). The size of the neighborhood context is 4 (exactly as shown in Figure 3).

MSRC dataset when the prior $\hat{P}(x|c)$ is empirically estimated directly from training data, instead of being obtained from the test set by statistical inversion. This result is remarkably close to that of Table 3(a), giving strong evidence of the power of the empirical Bayes approach to extract accurate contextual priors from unlabeled data.

3.3. Unsupervised image-level contexts

An important implementation issue is selecting the spatial support of the context. Figure 4 shows a typical example of how the prior changes as the context neighborhood becomes larger. Not surprisingly, as the size of the context is increased, the prior becomes smoother and more spatially uniform. It is especially interesting to consider the "limiting" case of priors that are constant over the entire image. To compute such image-level priors, we can take the following shortcut: instead of performing nearest-neighbor context search for each region in the image, we can simply declare that all the regions in the same image share the same context c and that $\hat{P}(y|c)$ for that context is the empirical distribution of observations in the image.

Table 3(c) shows the classification rates for image-level contexts obtained in this way. As compared to the performance of the neighborhood context shown in Table 3(a), we can see that the per-patch rate has increased by over 2%, and the per-class rate has increased by over 3%. Thus, for the MSRC dataset at least, the global or spatially constant contextual prior seems to be more effective than the neighborhood prior. But does this mean that our approach cannot derive any advantage from spatially varying neighborhood priors? The answer is a definite "no," for two reasons. First, as will be seen in Section 3.6, the neighborhood context can be more effective than the global context on image collections with different characteristics, such as the geometric context dataset [6]. A second, more compelling argument

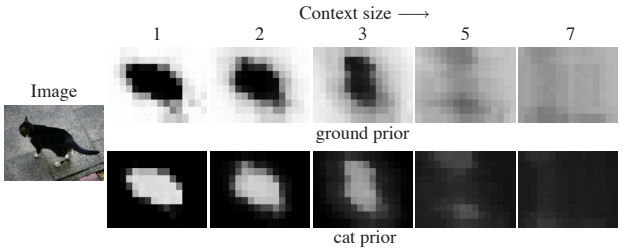

Figure 4. Dependence of contextual prior on context size (the number of concentric "rings" in Figure 3).

is that we can improve the overall system performance by combining neighborhood and image-level contexts, as will be described in the next section.

But first, we would like to discuss an intriguing connection that emerges between our proposed approach for computing image-level contexts and probabilistic Latent Semantic Analysis (pLSA) [5], a popular document model that has been successfully applied to images [15, 21]. This model was originally developed as an unsupervised procedure for discovering latent document structure in terms of underlying "topics" that generate the observed words. In relation to the framework of this paper, words correspond to observations y, and topics to class labels x. The distribution of words in a document d follows $\mathbb{P}(y|d) = \sum_x Q(y|x)\,\mathbb{P}(x|d)$, where $Q(y|x)$ is the conditional probability of word y given topic x and $\mathbb{P}(x|d)$ is a document-specific topic probability. In standard pLSA, both $Q(y|x)$ and $\mathbb{P}(x|d)$ are assumed unknown, and EM is used to estimate them simultaneously. However, if we assume Q to be known, then EM reduces to the optimization problem in eq. (2). This is known as the *fold-in heuristic*, and it is used to estimate the topic probabilities for new test documents after the likelihoods of words given topics are learned on a training set [5]. It has been argued that pLSA is not a "true" generative model because the prior over topics is conditioned on the "dummy" variable d [2]. But from the empirical Bayes perspective, d is actually the context, and pLSA can be thought of as a data-driven technique that uses the fact that a given group of words or observations all originated from the same document or image to infer a context-specific prior via statistical inversion.

It is very interesting that the fold-in heuristic for pLSA emerges as a special case of our approach, and that the empirical Bayes perspective helps to give a more satisfying interpretation of pLSA. Note, though, that in its full generality, our approach is quite different from pLSA. First, pLSA is a completely unsupervised procedure where the topic is a latent variable in a generative model and the objective is to maximize likelihood of the observed data. By contrast, our approach is discriminative and its objective is to minimize the average probability of error in a compound decision problem. Second, our approach is based on a much more general notion of context than pLSA. Our context is not restricted to image-level "dummy variables" but can vary from region to region in the same image, and regions from different images can share the same context. Furthermore, our approach offers the flexibility of combining image-level and spatially varying priors, as discussed next.

3.4. Combining neighborhood and image contexts

Intuitively, neighborhood and global contexts should be complementary, as they capture different types of dependencies. The neighborhood context can group patches from different images that are surrounded by similar patterns of

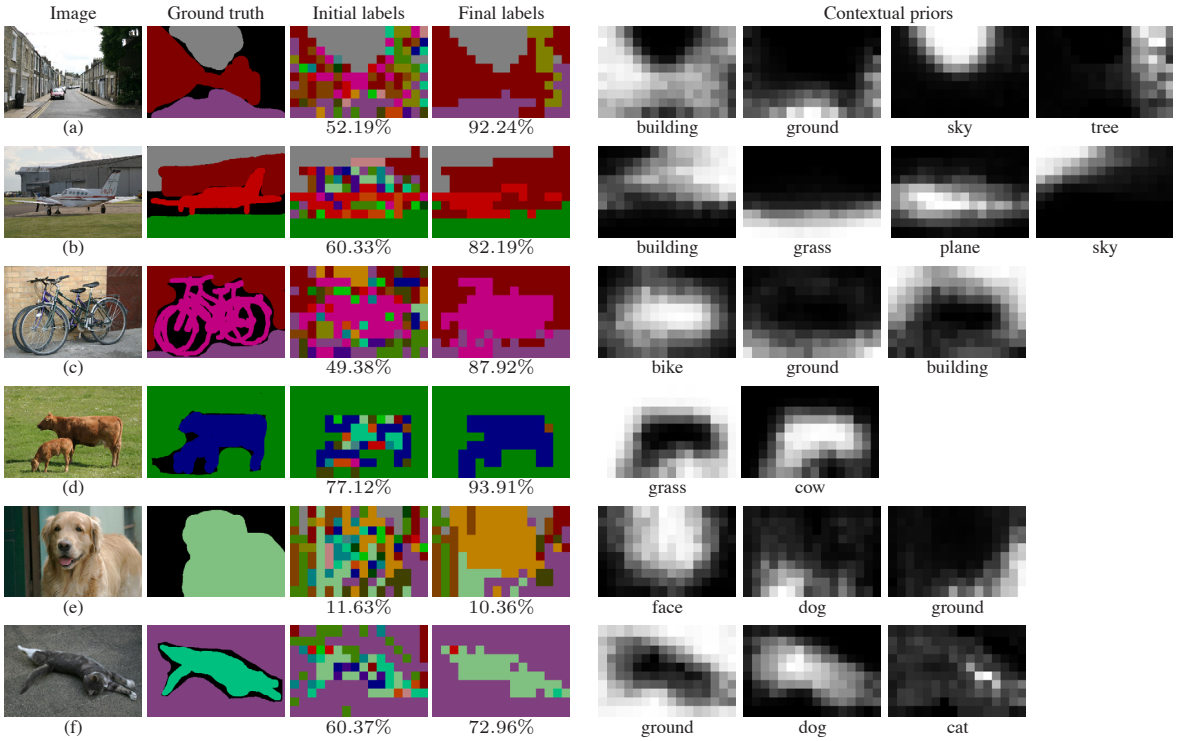

Figure 5. Examples of region classification on the MSRC dataset (best viewed in color). Each row shows (from left to right): the original image; the ground-truth labeling; the initial labeling produced by the local likelihood model (the percentage of correct labels is indicated below the image); the final labeling produced by the combined contextual model (Table 3(e)); contextual priors for the most common classes in the image.

observations, and can yield spatially varying priors. The image-level context can capture the dependence of all the patches within the image on the same underlying scene, but it can only produce priors that are constant over the entire image. How can we obtain priors that combine image-level regularization and spatial variability?

It is important to realize that the essential function of context within our empirical Bayes framework is to pool together image regions whose underlying class labels have the same expected statistical behavior. For this reason, the context representation should be designed to be as informative about the labels x as possible. In particular, if we have a class label sequence $\hat{\mathbf{x}}$ estimated with the help of some "oracle," we can try to define the context function in terms of this sequence instead of the raw observations $\mathbf{y}$. Accordingly, we take the class labels estimated with the help of image-level contexts as such an "oracle" and form a neighborhood context by histogramming these estimated labels using the concentric pattern of Figure 3. This procedure effectively combines image-level and neighborhood contexts by applying two rounds of label estimation in sequence, where the first round uses image-level contexts and its output is used to form neighborhood contexts in the second round. As can be seen from Table 3(d), this approach achieves an improvement of about 3% over the image-level contexts (c). Figure 5 shows the spatially varying priors computed with this approach on sample MSRC images.

Note that nothing stops us from iterating the above process further. Every time we compute an improved estimate of class labels over the images, we can re-compute the neighborhood context using the new labels and re-run the empirical Bayes algorithm to get a further improved estimate. This strategy, similar to iterated conditional modes [1], tends to converge very quickly (i.e., in about three iterations), and typically results in a further improvement of just under 1%.

3.5. Additional evaluation

The contextual priors described so far are extracted by statistical inversion directly from the unlabeled collection of test images, when all we know in advance is the local likelihood model $Q(y|x)$. As far as we know, this type of unsupervised context is completely novel. However, it is also true that by restricting ourselves solely to unsupervised context we are not utilizing all the possible information that may be present in the training set. In this section, we show that it may be possible to improve performance by incorporating additional supervisory information when it is available. Specifically, we show how to incorporate a supervised image-level prior obtained through discriminative training of global image models, as proposed by [20].

To learn a supervised image-level prior, we need a training sample of images labeled by the classes that they contain. Clearly, both patch-level and image-level labels can come from the same training set if it is labeled at the pixel level (as the MSRC dataset is). Alternatively, the image-level labels can come from a separate, weakly labeled col-

Method	Per-patch	Per-class
Local maximum likelihood	57.21	45.92
Neighborhood (unsupervised)	62.44	48.71
Image-level (unsupervised)	60.86	48.23
Neighborhood + image (unsupervised)	63.85	51.03
Unsupervised + supervised	**64.08**	**51.13**

Table 4. Results on the geometric context dataset.

lection. Given this additional supervisory information, we train binary support vector machines to predict the presence of each target class in the image as a whole. The image-level feature vectors on which the SVMs are trained are given by global histograms of quantizer labels for the SIFT, texton, and color features extracted from the image, concatenated with the histogram of per-patch class labels estimated using the unsupervised image-level prior. The output of the binary SVM corresponding to class x is converted to a probability in the standard fashion [10] and becomes a *supervised* image-level prior $P(x)$. This prior is then used to "modulate" the unsupervised contextual prior by multiplication, as suggested in [20]. Table 3(e) shows the performance obtained by the hybrid unsupervised/supervised system, which improves over the unsupervised priors by about 3%. This is our best performance, which exceeds [19], is comparable to [18, 21], but is below a few state-of-the-art supervised methods [11, 20]. Overall, this is very encouraging for a new method that is the first to explore an unsupervised notion of context. Unlike more mature supervised methods, ours can learn all it needs from a "pile" of labeled local patches, and recover a prior contextual model over classes "on the fly" from the unlabeled test set.

Most importantly, our results provide a convincing initial demonstration of the potential value of the empirical Bayes framework, confirming that it can extract contextual priors of non-trivial discriminative power from unlabeled data. The examples of Figure 5 show that the priors inferred by our method are perceptually plausible, in that they correctly capture the identities and overall spatial organization of the major classes in the image. In some cases, these priors can even be more accurate than the ground truth because they correctly "fill in" background regions where the labels are missing, such as the tree in (a). Many of the confusions of our method are also understandable: in (e), a close-up of a flesh-colored dog's face is confused with a human face, and in (f), a cat is confused with a dog. Finally, Figure 6(a) shows the magnitude of improvement of the contextual model over the local likelihood model for individual images. The average per-image improvement is 18.46%. This is the extent to which we can successfully "denoise" the image observations to recover the original class labels.

3.6. The geometric context dataset

In this section, we report results of our method on the 300-image geometric context dataset [6]. The seven classes in this dataset correspond to geometric surface types: sky, ground, solid, porous, and vertical facing left, right, and center. We use five-fold cross-validation with the same subsets as [6], and the results are given in Table 4. Unlike on the MSRC dataset, the image-level context here is weaker than the neighborhood context, and incorporating additional supervisory information (last line of the table) gives little improvement. This is due to the fact that the relative frequencies of the geometric classes are much more uniform than those of the MSRC classes. Also, while a typical MSRC image contains only a small subset of the 21 classes, a typical geometric context image contains a majority of the seven geometric classes, resulting in "flatter" image-level priors.

(a)

	sky	ground	vert.	sky	ground	vert.
				Our method		
sky	92.5	0.3	7.2	90	0	10
ground	1.4	78.0	20.7	0	78	22
vertical	3.9	11.8	84.3	2	9	89

(b)

	center	left	right	solid	porous	center	left	right	solid	porous
						Our method				
						Hoiem et al. [6]				
center	28.1	7.8	17.0	29.4	17.6	55	2	6	18	19
left	14.6	18.6	21.9	30.7	14.0	46	15	4	21	15
right	16.4	7.4	35.6	24.8	15.8	38	3	21	21	17
solid	11.3	2.8	5.1	64.9	15.8	20	2	3	50	26
porous	5.8	2.4	3.3	13.8	74.7	14	1	2	8	76

Table 5. Comparison with [6] for (a) the three main classes and (b) the five vertical sub-classes.

Table 5 compares the performance of our system with that of Hoiem et al. [6], which is specifically tailored for the task of geometric context classification. The two approaches have comparable performance on the three major classes (sky, ground, and vertical), but our approach generates more confusion between the sub-classes based on vertical surface orientation (left, right, center). We conjecture that specialized non-local cues used by [6], such as statistics of lines and vanishing points, are necessary to achieve higher accuracy on those classes. Figure 7 gives a few example images classified by our system. Example (d) shows that, as with many other image labeling approaches, the addition of a prior can sometimes make the initial labeling worse by converting a large chunk of the image to an incorrect class. Figure 6(b) confirms that there is a tendency for contextual classification to decrease the accuracy of local labelings that are poor to begin with. Nevertheless, the application of our contextual model to this dataset results in an average improvement of 6.8% per image.

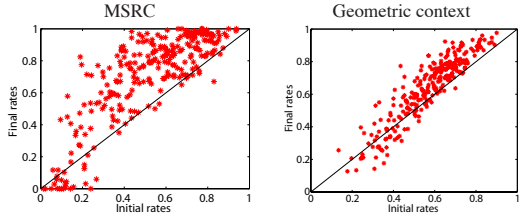

Figure 6. Final (contextual) vs. initial (local) performance for individual images. All the data points above the diagonal line correspond to images whose classification rates have improved. For the MSRC dataset, the average (resp. maximum) improvement is 18.46% (resp. 63.51%), and for the geometric context dataset, it is 6.81% (resp. 27.33%).

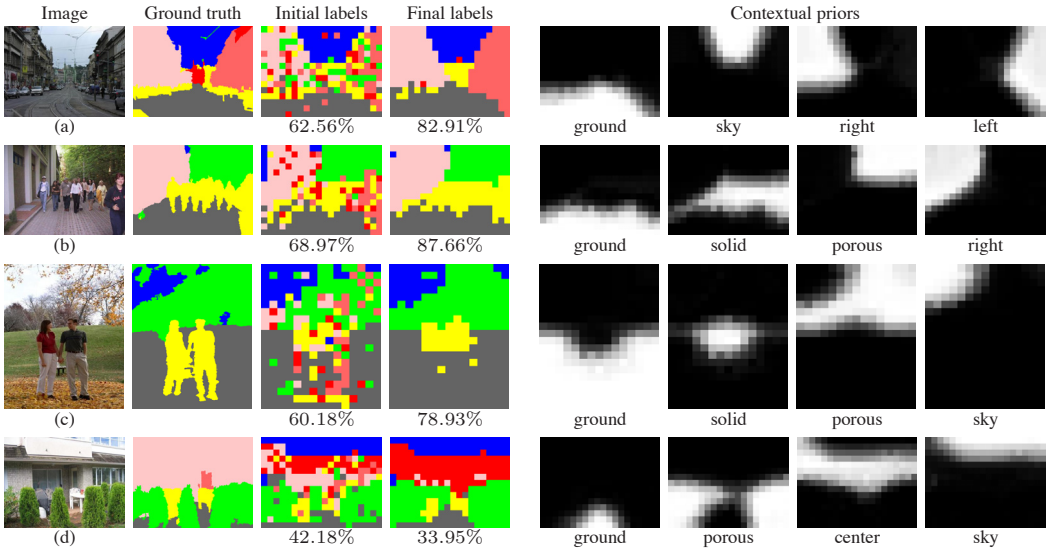

Figure 7. Examples of region classification on the geometric context dataset (best viewed in color).

4. Conclusion

In this paper, we have presented a solution to the region classification problem based on the empirical Bayes tradition [13] and on recent developments in information-theoretic denoising [22]. We argue that at the scale of an individual patch, the stochastic mapping from a semantic class label to the corresponding image observation can be learned reasonably well from a sufficiently representative sample of labeled patches, while the detailed class composition of an image may vary significantly from scene to scene, and is rather more difficult to learn in advance. Formalizing this intuition results in a novel, principled, and simple method for extracting contextual information directly from unlabeled test data, which can potentially lead to advances in labeling of large-scale, sparsely labeled datasets such as that of [3]. To make our method scale to such datasets, we plan to replace the exhaustive search that is currently used to estimate contextual probabilities by fast approximate search techniques such as locality sensitive hashing [16] or randomized forests [18, 20]. Another avenue for future work is to make our method completely unsupervised by using EM (as in pLSA) to simultaneously infer the local likelihood model and the contextual priors.

Acknowledgments. This research was supported in part by NSF grant CRI 0751187. The authors are also grateful to Jean Ponce and Alyosha Efros for helpful discussions.

References

[1] J. Besag. On the statistical analysis of dirty pictures. *J. Roy. Stat. Soc. B* 48, pp. 259-302, 1986.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *JMLR* 3, pp. 993-1022, 2003.

[3] S. Divvala, A. Efros, and M. Hebert. Can Similar Scenes help Surface Layout Estimation? *Internet Vision Workshop, CVPR* 2008.

[4] G. Heitz and D. Koller. Learning Spatial Context: Using Stuff to Find Things. *ECCV* 2008.

[5] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. SIGIR*, 1999.

[6] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. *ICCV* 2005.

[7] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV* 43(1):29-44, 2001.

[8] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Tokyo: Springer-Verlag, 2001.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV* 60(2):91-110, 2004.

[10] J. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, MIT Press, 1999.

[11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie. Objects in context. *ICCV* 2007.

[12] H. Robbins. Asymptotically subminimax solutions of compound decision problems. In *Proc. 2nd Berkeley Symp. Math. Statist. Probab.*, vol. 1, pp. 131–148, 1951.

[13] H. Robbins. An empirical Bayes approach to statistics. In *Proc. 3rd Berkeley Symp. Math. Statist. Probab.*, vol. 1, pp. 157–163, 1956.

[14] C.P. Robert. *The Bayesian Choice*, 2nd ed. Springer-Verlag, New York (2001).

[15] J. Sivic, B. Russell, A. Efros, A. Zisserman, W. Freeman. Discovering objects and their location in images. *ICCV* 2005.

[16] G. Shakhnarovich, T. Darrell, and P. Indyk, eds. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, 2006.

[17] C. Schmid. Constructing models for content-based image retrieval. *CVPR* 2001, vol. 2, pp. 39-45.

[18] F. Schroff, A. Criminisi, and A. Zisserman. Object Class Segmentation using Random Forests. *BMVC* 2008.

[19] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV* 2006.

[20] J. Shotton, M. Johnson, and R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. *CVPR* 2008.

[21] J. Verbeek and B. Triggs. Region classification with Markov field aspect models. *CVPR* 2007.

[22] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, M.J. Weinberger. Universal discrete denoising: known channel. *IEEE Trans. Info. Theory* 51 (1): 5–28, 2005.