Low-level Hierarchical Multiscale Segmentation Statistics of Natural Images

Emre Akbas, Narendra Ahuja, *Fellow, IEEE,*

Abstract—This paper is aimed at obtaining the statistics as a probabilistic model pertaining to the geometric, topological and photometric structure of natural images. The image structure is represented by its segmentation graph derived from the low-level hierarchical multiscale image segmentation. We first estimate the statistics of a number of segmentation graph properties from a large number of images. Our estimates confirm some findings reported in the past work, as well as provide some new ones. We then obtain a Markov random field based model of the segmentation graph which subsumes the observed statistics. To demonstrate the value of the model and the statistics, we show how its use as a prior impacts three applications: image classification, semantic image segmentation and object detection.

Index Terms—Natural image statistics, low-level hierarchical segmentation, Markov random field.

1 Introduction and Related Work

A natural image is far from being a random configuration of pixels. Rather, it exhibits a high degree of organization, e.g., reflected in its spatial and spectral properties, such as geometric, photometric and layout properties. The work on natural image statistics attempts to derive probabilistic models of image properties. The most widely accepted definition of a natural image is that it has a commonly encountered statistical structure to which the human visual system has adapted [19].

Natural image statistics have received significant attention in recent years. Models developed for it have been used in numerous applications such as denoising [28], [40], inpainting [28], scene categorization [36], contextual priming for object detection [35], sensory processing in biology [29], saliency detection [42] and texture synthesis [17] among many others [33].

Previous work on natural image statistics can be broadly grouped in two categories. The work in the first category analyzes natural images and seeks to obtain properties, laws, invariances, etc. characteristic of natural images. Some findings of this work are the following: (1) the image spectra obey a power law (see [29], [36], [33] for a list of references), (2) scale invariance of statistics [29], [43], [33], (3) responses of image patches to certain filters follow non-Gaussian, highly kurtotic, heavy-tailed distributions [18], [28], [33], (4) edges are dominantly oriented either horizontally or vertically [10].

The work in the second category is aimed at developing models of natural image statistics for use in various image processing and computer vision applications. A prominent and widely adopted model in this category is the Field of Experts (FoE) model [27], [28] proposed by Roth and Black. FoE learns a set of linear filters whose responses are modeled by heavy-tailed, kurtotic distributions within a high-order Markov random field model. They successfully demonstrate the use of FoE in denoising and inpainting applications. Weiss and Freeman proposed a more efficient learning algorithm for the FoE model in [40]. FoE has been further improved by Heess et al. to allow for bimodal potentials and they used their model for texture synthesis. In [9], Cho et al. questioned the heavy-tailed distribution assumption and proposed a new image prior that adapts itself to the content, i.e. the type of underlying texture. Torralba exploited statistics of natural images for scene categorization [36] and for contextual priming for object detection, i.e. learning priors on possible locations and sizes of objects. We refer the reader to [33] for a more complete list of other applications of natural image statistics.

Both of the aforementioned categories have one thing in common: all perform analysis that are pixel, patch or subband-based. Limited work has been done on the statistics of image features such as contours and regions. Alvarez et al. [5] analyzes the size distribution of image regions obtained using an intensity-based segmentation into connected components. Ren and Malik [26] model the statistics of boundary contours and derive a prior model for contour shape. The "dead leaves" model [33] uses a low-level occlusion process to estimate the statistics of image partitions. Ghosh et al. [15], [14] uses nonparametric Bayesian processes to model image partitions. Carreira and Sminchisescu [7] have investigated figure-ground segmentation based on mid-level properties of combinations of image regions. A wide range of statistics of high-level and human segmentations of images have also been reported [22], [30]. However, to the best of our knowledge, statistics of properties that capture geometric and topological structure of images, e.g., as defined by regions obtained by low-level image

- *E. Akbas is with the Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA, 93106. He was with the Beckman Institute for Advanced Science and Technology at the University of Illinois while conducting this research.*
- *N. Ahuja is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61820.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2299809, IEEE Transactions on Pattern Analysis and Machine Intelligence

segmentation do not appear to have been investigated. This paper is aimed at computing such statistics of natural images.

It has been shown that low level segmentation serves as a useful seed for simultaneous discovery, modeling, recognition and segmentation of objects in images [34]. Therefore, the performance of the segmentation based algorithms for recognition, etc., may be improved by using segmentation related statistics of natural images, e.g., as priors in a Bayesian framework. For a variety of reasons, including the dependence of the available segmentation algorithms on user provided parameters, segmentation statistics do not appear to have been obtained for natural images.

1.1 Overview of our approach

In this paper, we compile as well as use the segmentation statistics from a large number of natural images. First, we present the statistics we choose to estimate and discuss the reasons for choosing them. Next, we present a probabilistic model to define and learn these statistics. Since our low level segmentation is represented by a graph, our estimation procedure involves estimating statistics of selected properties of this graph. We use Markov random field (MRF) based modeling for this purpose. The statistics we have estimated confirm some of the previous findings, included in the past work (e.g. dominant orientations are horizontal and vertical). They also yield new findings, in terms of the properties of segmentation, and therefore outside the scope of the previous work (e.g. number of regions versus photometric scale follows an exponential distribution)

Next, to demonstrate the value of the statistics, we use them as priors in three applications: image classification, semantic image segmentation and object detection. It is expected that, as usual, the use of priors would improve the performance of the algorithms. This expectation is borne out by our experimental results.

We use the low-level multi-scale segmentation algorithm of [3] in our experiments. The reason we specifically chose this algorithm is that it (i) does not require any major user supplied parameters while it provides all, previously unknown, naturally occurring segmentations, which are perceptually valid and organized in a hierarchy [1], and (ii) it has been shown to outperform other available algorithms on a low-level segmentation benchmark [3].

There are three main contributions of this work: (1) we present some potentially useful statistics of low-level segmentation of natural images for the first time, and (2) we present a probabilistic model of segmentation which we use to learn these statistics, and (3) we demonstrate the use of this model in three applications.

2 Models and Statistics

In this section, we first briefly describe the segmentation graph, the representation we use for the multi-scale image segmentation, and its properties we use to capture the underlying image. Next, we provide some statistics of these properties over a large number of natural images. Finally, we present the proposed Markov random field (MRF) based model, and describe how to learn it from a set of given images.

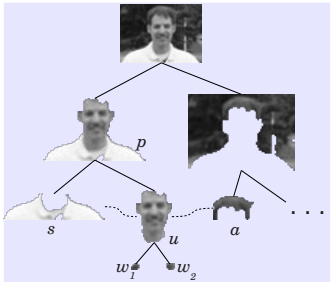

Fig. 1. A sample segmentation tree. For a given region (or, to be interchangeably referred by its associated node) u, the segmentation tree relates it to its parent node p and its children nodes w_1 and w_2. The connected segmentation tree of [2] adds edges between u and its sibling s. Our segmentation graph, in addition, includes edges between u and its adjacent regions, i.e., regions sharing border with u. In the figure, a is one such region. The edges to the sibling and adjacent regions/nodes are shown dashed. By adding the dashed edges to the segmentation tree, we obtain the segmentation graph.

2.1 Segmentation graph and region properties

The segmentation algorithm in [3] partitions a given image into homogeneous regions of a priori unknown shape, size and degree of photometric homogeneity. The algorithm organizes all detected regions into a hierarchical structure called the segmentation tree [1] which captures the low-level, spatial, and photometric image structure in a hierarchical manner. Nodes at upper levels correspond to larger segments, while their children nodes capture smaller embedded details. Fig. 1 shows that a segmentation tree includes edges between nodes corresponding to a region and those corresponding to smaller regions contained within the region, and a connected segmentation tree of [2] includes additional edges to sibling nodes, i.e., corresponding to those regions that "surround" the region. The "surround" is defined in [2] by introducing neighboring Voronoi regions, analogous to the conventional point Voronoi neighbors in a point pattern. Our new structure, segmentation graph, in addition includes edges to the extreme case of adjacent regions, i.e., regions at zero distance, sharing border.

We describe each node, i.e. region in the segmentation graph by a set of its intrinsic (geometric and photometric) properties, as well as relative properties (capturing region topology and layout). A property is intrinsic if it represents only the region itself. A relative property, on the other hand, describes how the region is related to its parent and regions that are its siblings or other (Voronoi or adjacent) lateral neighbors, linked to it in the segmentation graph.

Choosing the specific properties: Conventional image properties are typically measured on pixels or windows of pixels. Therefore, they do not capture interesting geometric or topological aspects of image structures; they are limited to estimating properties of sets of pixel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2299809, IEEE Transactions on Pattern Analysis and Machine Intelligence

colors or simple geometric properties such as of edges. The variety of geometric, photometric and topological structure captured by the segmentation graph, and therefore the ability to measure related specific properties provides image properties not available through the usual approaches. The power of region based representation of images, and particularly, of properties of the types described below, has been demonstrated by prior empirical results, wherein different combinations of properties have been shown to be effective in solving a range of problems [13], [4], [34].

Below we list the properties we have used. It may be a redundant set. Analysis and experimentation in future work may help refine these properties into a more compact set, which may more effectively represent semantic and low level contents of images.

Intrinsic properties: The intrinsic geometric properties we use are concerned with the following aspects of a region: (1) Area (normalized by the image area). (2) Squared perimeter over area, which indicates how compact the region is, the most compact shape being the disk. (3) Eccentricity, i.e. scalar that specifies the eccentricity of the ellipse that has the same second-moments as the region. (4) The first four Hu moment invariants. (5) Orientation. (6) Perimeter. (7) Solidity, i.e. the proportion of the pixels in the convex hull that are also in the region. (8) Extent, i.e. the ratio of pixels in the region to pixels in the bounding box of the region. (9) Major and minor axes lengths normalized by image perimeter. (10) Location of the center of mass w.r.t. the image coordinates. (11) A pyramid histogram of oriented gradients (PHOG) where we evenly divide the bounding box of the region into 1, 4, and 16 cells, thereby obtaining 3 levels.

With respect to the intrinsic photometric properties, the grayscale distribution within a region is represented by mean and standard deviation of the region's intensities. For color images, we use mean and standard deviation of each of the RGB channels.

Relative properties: These include the following geometric, photometric as well as topological properties characterizing a region's relationships with other regions. The geometric properties include: (1) Outerring area, i.e. area of the region minus the total contained area of its children, divided by the total region area (ring + children areas). (2) Normalized area, i.e. area divided by the parent's area. (3) Location of the center of mass w.r.t. the parent's coordinate system.

The photometric properties we use are as follows. (1) Each region is associated with a photometric scale which is the contrast level at which it is detected during segmentation. (2) Mean and standard deviation of the contrast along its boundary,

Following are the topological properties of a region that we use. (1) Number of children. (2) Area variance of the children. (3) Sibling-context vector, i.e. a vector which records the general direction in which the region sees its siblings located. We refer the reader to [34] for the details on above properties.

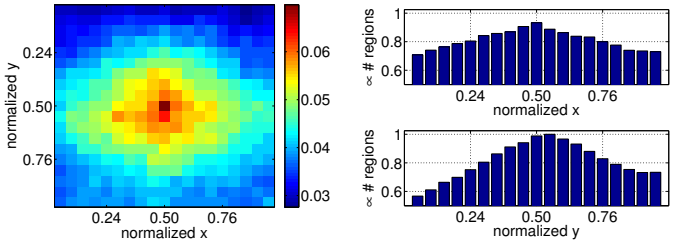

Fig. 2. (Left) 2D histogram of xy-coordinates of center-of-mass of regions. For a given bin, color encodes the (relative) number of regions whose center of mass fall in that bin. (Right) 1D histograms of x and y coordinates.

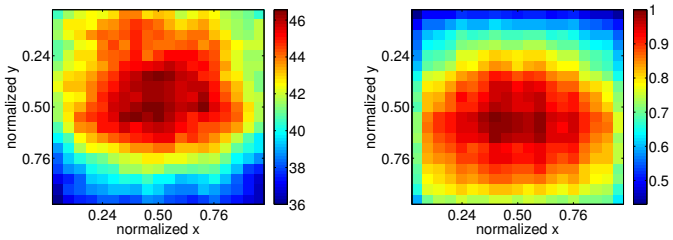

Fig. 3. (Left) Histogram of boundary contrasts. For a given bin, color encodes the average contrast of region boundaries falling in that bin. (Right) Histogram of Canny edges. For a given bin, color encodes the (relative) number of edges falling in that bin.

A different set of relative properties, which emphasize appearance modeling based on texton and brightness histograms, were proposed in [7]. While these properties might be useful for figure-ground segmentation, we have not included these properties in our analysis for two reasons: (i) they do not make use of the topological structure of regions and (ii) [3] produces mostly homogeneous regions, describing their appearance using texton and brightness histograms would bring little additional benefit.

2.2 Statistics of selected properties

We randomly collected a set of 2000 images[1]. We segmented each image and obtained its segmentation graph along with the property vectors described in the previous section. Below we provide sample statistics (in the form of histograms) of the following properties: (1) location of center of mass, (2) orientation, (3) number of regions versus the photometric scale, (4) area, and (5) depth and average branching factor of the segmentation tree. Among the large set of properties presented in the previous section, we found these properties to be the most interesting ones in that their statistics reveal potentially useful interpretations about natural images.

2.2.1 Location of center of mass

Fig. 2 gives the histograms of normalized row (y) and column (x) coordinates of the center of mass of regions. Top left corner of the image is assumed to be $(0, 0)$ and bottom-right is $(1, 1)$. The histograms suggest that there are more regions around the center of the image

1. 600 images from the "Flickr random image generator" website (http://beesbuzz.biz/crap/flrig.cgi) and 1400 images from PASCAL VOC 2010 dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2299809, IEEE Transactions on Pattern Analysis and Machine Intelligence

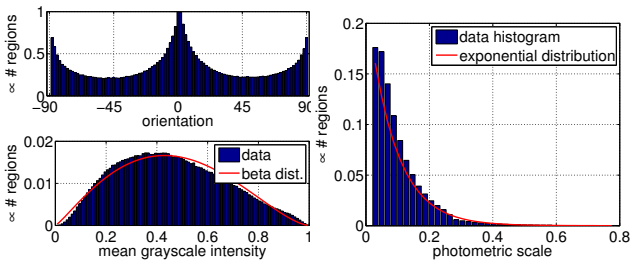

Fig. 4. (Left,top) Histogram of region orientations. (Left,bottom) Histogram of mean grayscale intensity of regions. (Right) Histogram of photometric scales of regions.

than the surround. The number of regions decrease in any direction from the center to image borders. This could be explained with the photographer bias [37], a natural tendency of photographers to place the object of interest near the center of their field of view. Because the focus is more likely around the center, the edges there are sharper, i.e. their ramp widths are narrower, than the edges at the periphery of the center. Out of focus blur causes region boundaries, hence regions, in the periphery die earlier as the photometric scale is increased during segmentation. To quantify this effect, we computed a spatial histogram of boundary contrasts (Fig. 3 (left)) where for each bin we computed the average contrast of the region boundaries spatially falling in that bin. One can see that the average boundary contrast is highest around the center of the image and decreases towards the image borders.

In Fig. 2 (right), we see that the distribution of x seems to have no bias for either the left or the right half of the image. However, y coordinates seems to be biased towards the $[0.5, 1]$ interval, which means more regions have centers of mass in the lower halves of the images than in the upper halves. Specifically, there are about 11% more[2] regions in the lower halves, while the difference between left and right halves is slightly less than 1%. The bias in the distribution of y coordinates might be attributed to the natural existence of an omnipresent ground, with many more objects located on it than at distances "above the ground." This effect seems to more than compensate for the reduction in the number of regions in the lower half that may be expected due to reduction in the average boundary contrast in the lower part (Fig. 3 (left)), which, in turn, is presumably a manifestation of light coming from above. Our observation of a larger number of regions lower in the image is in accord with previous work [39], [6] reporting that there is more texture/detail in the lower part of natural images.

2.2.2 Orientation

Fig. 4(Left,top) gives the histogram of region orientations. We represent region orientation by the angle between the x-axis of the image and the major axis of the ellipse that has the same second-moments as the region. The shape of the histogram shows that there are two dominant orientations: horizontal ($0°$), and vertical ($-90°$ and $90°$). It has previously been reported [10], [36] that the distribution of edge orientations in natural images is biased towards horizontal and vertical directions, which is consistent with our finding here. Presumably, this bias could be attributed to the existence of an omnipresent ground and that objects counter the gravitational force more efficiently when standing at horizontal and vertical directions than oblique angles.

2. Percent difference between two values a and b is calculated as $100|a-b|/(0.5(a+b))$.

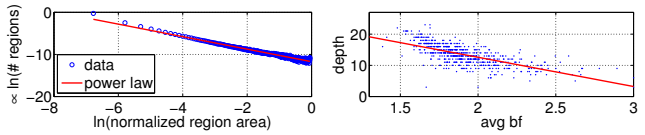

Fig. 5. (Left) Log-log plot of number of regions as a function of their area. (Right) Depth and average branching factor (ABF) of segmentation trees. Each point represents a tree.

2.2.3 Photometric scale

Fig. 4 (Right) shows the histogram of the photometric scale of the regions detected by our segmentation algorithm. Other than the decrease, as expected, in the number of regions as the lowest acceptable contrast for the regions is increased, the histogram quantifies the decrease: the number of regions versus the photometric scale follows an exponential distribution (tested using Pearson's χ^2 test). It appears that this exponential trend can be explained by a simple model: suppose that a region X is surrounded by N other regions $Y_1, Y_2, \ldots, Y_N$ which are spatially adjacent to it. Let G_X represent the grayscale intensity of region X, and assume that all the pixels within a region have the same intensity. Then, the photometric scale of region X can be written as $\sigma_X = \min_i(|G_X - G_{Y_i}|)$ [1], [3]. The distribution of mean grayscale intensity of regions is given in Fig. 4 (Left,bottom). While the histogram looks like a truncated Gaussian distribution, the data fits better to a beta distribution. Therefore, G_X, G_{Y_i} can be taken as beta-distributed random variables. We compute the distribution of σ_X by simulation, as we are not aware of a closed-form expression for it. We draw a large sample of values for G_X, G_{Y_i} from the fit beta distribution, and compute σ_X. It turns out that this simulated σ_X also follows the exponential distribution.

2.2.4 Area

Fig. 5 (Left) shows the log-log plot of (relative) number of regions as a function of normalized region area. As the plot suggests, most of the regions are small. In fact, the percentage of regions which have area less than 5% of the image is 98.81%. Region areas seem to follow a power law (confirmed using a χ^2 test) shown in red. This result seems to be a property of sizes of objects (and object parts) projected onto the image plane. Region area depends on the viewing distance and the apparent size of objects (and their parts) as captured by segmentation. Assuming the area statistics subsume all possible viewing angles and distances, it is reasonable

to speculate that object sizes in nature follow a power law. However, we are not aware of any work on such statistics of objects.

2.2.5 Depth and average branching factor

Both the depth and average branching factor (ABF) of segmentation trees seem to follow the normal distribution with $\mu = 13.34, \sigma = 3.31$ and $\mu = 1.92, \sigma = 0.25$, respectively. One interesting observation on depth and ABF values is that they are inversely proportional (see Fig. 5 (Right), line fit to data has negative slope), i.e. more branching less depth, and vice versa. This implies a "conservation of number of regions" rule.

2.3 MRF modeling of the segmentation graph

Let S be the segmentation graph of an image I. Natural image statistics can be easily specified if we have a model for $p(I)$, i.e. the probability that I is an image. In this work, we develop this model in terms of the segmentation graph properties of the image by assuming the probability I is equivalent to the probability of observing its segmentation graph S, i.e. $p(S) \simeq p(I)$.

We assume that a node in the segmentation graph together with its immediate neighbors (parent, children, sibling (Voronoi or adjacent) regions) form a maximal clique of the segmentation graph[3]. Let r_k denote the property vector extracted from the clique defined with reference to the k^{th} region. r_k is the concatenation of the intrinsic properties of the node and its relative properties (described in Sec. 2.1) determined through the clique. Suppose S has K nodes, then our proposed model is:

$$p(S;\Theta) = \frac{1}{Z(\Theta)} \prod_{k=1}^{K} \psi(r_k;\Theta), \tag{1}$$

where we parametrized $p(S)$ with our model parameters Θ and $\psi(\cdot)$ is the clique-potential function, and $Z(\cdot)$ is a normalization factor called the partition function. Using the Markov-Gibbs equivalence, we can write (1) as:

$$p(S;\Theta) = \frac{1}{Z(\Theta)} e^{-E(S;\Theta)} \text{ where } Z(\Theta) = \int_V e^{-E(s;\Theta)} ds, \tag{2}$$

where V is the space of all possible segmentations and $E(\cdot)$ is the energy function which is defined as:

$$E(S;\Theta) = \sum_{k=1}^{K} U(r_k;\Theta). \tag{3}$$

$U(\cdot)$ is called the log-potential function, in fact, we have $U(\cdot) = ln\ \psi(\cdot)$. We discuss the form $U(\cdot)$ in Section 2.3.2.

2.3.1 Estimating the parameters, Θ, of the model

It is well known that maximum likelihood (ML) estimation of Θ is intractable because of the partition function $Z(\Theta)$. A simple approximate scheme to ML estimation is the pseudo-likelihood estimation which is an asymptotically consistent estimator of ML [21]. Pseudo-likelihood ($\mathcal{PL}$) approximation is based on the conditional probability of a node given its immediate neighbors. In our case, we define it as the probability of observing r_k:

$$p(r_k;\Theta) = e^{-U(r_k;\Theta)} \Big/ \int_\Omega e^{-U(r;\Theta)} dr \tag{4}$$

where Ω is the space of all possible values that a property vector can take. Then, $\mathcal{PL}$ is:

$$\mathcal{PL}(\Theta) = \prod_{k=1}^{K} p(r_k;\Theta) = \prod_{k=1}^{K} \frac{e^{-U(r_k;\Theta)}}{\int_\Omega e^{-U(r;\Theta)} dr}. \tag{5}$$

One way to maximize $\mathcal{PL}$ is by gradient ascent where the gradient is:

$$\frac{\partial ln(\mathcal{PL}(\Theta))}{\partial \Theta} = K \frac{\int_\Omega e^{-U(r;\Theta)} \frac{\partial U(r;\Theta)}{\partial \Theta} dr}{\int_\Omega e^{-U(r;\Theta)} dr} - \sum_{k=1}^{K} \frac{\partial U(r_k;\Theta)}{\partial \Theta} \tag{6}$$

Note the integrals in the first term above. Depending on the dimensionality of the property vectors, numerical integration techniques ranging from simple quadrature methods to Monte Carlo methods could be used. Other than gradient ascent, a recently popular method to optimize Eq. (2) is the contrastive-divergence method which has been used in the FoE models. We do not use this method because it would require us to draw samples from V, the space of all possible segmentations.

2.3.2 Potential functions

The criterion for a good potential function, $U(\cdot)$, is such that it should be minimized when it describes a clique, i.e. a node and its immediate neighbors as completely as possible. That is, $U(r;\Theta)$ should be minimum, if the property vector r is describing a valid region together with the regions corresponding to its segmentation graph neighbors. A natural choice would be to use a negated probability density function (pdf). We note that when $U(r;\Theta)$ is in the form:

$$U(r) = -\ln(q(r)) \tag{7}$$

where $q(\cdot)$ is a pdf, the maximization $\mathcal{PL}$ becomes much easier due to the fact that the denominator in (4) always evaluates to 1 since $q(\cdot)$ is a pdf. This saves a lot of computation because we no longer need the first term and the high-dimensional integrals in the gradient (6).

Following the observation above, we choose $q(\cdot)$ to be a Gaussian mixture model (GMM) due to its generality and some empirical evidence that GMMs model region properties well in certain tasks [4]. With these choices, maximizing pseudo-likelihood boils down to fitting a GMM, for which we use the standard expectation-maximization algorithm.

3 Experiments

In this section, we demonstrate the use of our proposed model in three different applications: image classification, semantic segmentation and object detection.

3.1 Image classification

We make use of our proposed model for image classification using the Fisher kernel approach [20], [24]. Let S_1 and S_2 be two different segmentation graphs which are generated by the prior $p(S;\Theta)$. Then, a representative feature vector for S_1 is $f_1 = \nabla_\Theta \log p(S_1;\Theta)$. f_1

3. Similarly, in the Fields-of-Experts model, a 5x5 square patch centered on a pixel is assumed to define a maximal clique.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2299809, IEEE Transactions on Pattern Analysis and Machine Intelligence

represents S_1 because the gradient of the prior model evaluated at S_1 describes the directions in which the model parameters should be changed to best fit S_1. In fact, f_1 is a sufficient statistics for S_1 [20]. Similarity between between S_1 and S_2 is given by the kernel $K(S_1, S_2) = f_1^T F^{-1} f_2$, where F is the Fisher information matrix of $p(S;\Theta)$. Appropriately scaled f_1 (i.e. $F^{-\frac{1}{2}} f_1$) is called the Fisher vector representing S_1 (see [24], [25] for further details). Note that the size of the Fisher vector depends only on the number of parameters of $p(S;\Theta)$ and not on the size (number of nodes) of S_1. Efficient linear classifiers can be trained on Fisher vectors.

We used the PASCAL VOC 2007 dataset. In our experiments, we used all the features described in Sec. 2.1 with the exception that we compressed PHOG features using PCA due to the high dimensionality of the PHOG vectors (21 cells, each 8 dimensional, hence 168 dimension in total). Instead of fixing the reduced number of PHOG dimensions, we allowed it to vary as a parameter of the system. In addition to this parameter, the only parameter of our model is the number of Gaussian components in the potential function (Eq (7)). We tuned these two parameters by cross-validation on the "trainval" set.

First, we trained a single universal prior (as described in Sec. 2.3.1) using all the images of *all classes*. Then, we extracted the Fisher vectors using this prior and trained linear Support Vector Machine (SVM) classifiers [8]. We call this universal prior model as SS-U, short for "universal segmentation statistics".

While experimenting with SS-U, we observed that the optimal parameters, i.e. # of Gaussian components, and the reduced dimensionality of PHOG, are different for different classes. Following this observation, we trained another system where we learned a separate prior model for *each class*. We call this system as SS-PC, short for "segmentation statistics per class".

The average APs obtained by SS-U and SS-PC are shown in Table 1, together with the state-of-the-art results on this dataset. Although our proposed models does not perform as well as the best method available, SS-PC seems to be among the top performing methods. All the methods above SS-PC use sophisticated and costly learning algorithms. "Iterative Contextualizing" [32] iteratively combines object detection and image classification. Similarly "Cls + Loc" [16] combines the best method of PASCAL VOC 2007 with a costly sliding window-based object detector. Multiple kernel learning (MKL) [41] also trains a costly learning system and uses thousands of features. The "kernel codebook" [38] and the Improved Fisher Kernel (IFK) [25] are comparable to our proposed model in simplicity but we are working with less features (per image) than they do. On average, a typical 500x500 image gives us only a few hundred regions (the actual average over the 2000 image dataset is about 350), whereas [25] extracts around 5000 features (every 16 pixels, at five scales). Another difference between SS-PC and [25] is that instead of using the "spatial pyramid"[25] approach, we store the location information within the feature vector. This is a simpler approach than the spatial pyramid where one needs to train separate models per "spatial cell". The performance difference between SS-PC and [25] could be attributed to both different features (i.e. regions vs dense patches) and spatial aggregation schemes.

Method	AP	Method	AP
Iterative Contextual. [32]	70.5	Best of VOC07 [11]	59.4
Cls + Loc [16]	63.5	IFK (SIFT) [25]	58.3
MKL [41]	62.2	SS_U	53.4
SS-PC	61.1	Standard FK [25]	47.9
Kernel Codebook [38]	60.5		

TABLE 1
Comparison of the proposed models SS-U and SS-PC with the state-of-the-art on PASCAL VOC 2007.

3.2 Semantic Segmentation

In our second experiment, we show how to use image priors to help improve semantic segmentation of images. To this end, we use the MSRC-21 dataset which contains 591 images of 21 classes. The task is to label every pixel of the test images with one of these 21 labels. We follow the practice of [31] and randomly split the set into 276 training and 315 testing images.

Our semantic segmentation model is as follows. Suppose that we have a classifier which can predict the semantic label of a region. To label a pixel, we first classify the regions that contain it. Note that a certain pixel might be included in more than one region because of the segmentation hierarchy. The classifier returns its predictions along with confidence scores, or probabilities. Then, we choose the label with the maximum confidence and assign it to the pixel in question. We use a SVM with RBF kernel as our region classifier and train it on all the regions of all the training images.

Now, let us look at how image priors help this process. The region classifier is actually trained to compute the probability $p(c|r)$ where c is the class variable and r represents a region, i.e. its properties. However, we are also given the images that contain these training regions. In fact, we can learn $p(c|r,S)$[4] instead of just $p(c|r)$, where S is the segmentation graph of the image that contains region r. If we write out:

$$p(c|r,S) = \frac{p(r,S|c)p(c)}{p(r,S)} = \frac{p(r|c)p(S|c)p(c)}{p(r,S)} \quad (8)$$

where we assumed conditional independence of r and S given the class c. Then,

$$p(c|r,S) \propto p(c|r)p(S|c). \quad (9)$$

The first term on the right hand-side above is the region classifier and the second term is the class-conditional image prior. This expression shows the contribution of the image prior in region classification.

Implementation of the prior: From a theoretical point of view, one can learn a prior $p(S;\Theta)$ (as described in Sec. 2.3.1) using only those images that contain objects

4. This must actually be $p(c|r, S\backslash r)$ but we can assume $S\backslash r \approx S$ as the average region size is small compared to the size of the image.

of class c, and then use it to compute $p(S|c)$.

There are two problems with this. First, one needs to compute the value of the partition function, $Z(\Theta)$, but this is intractable. Second, $p(S;\Theta)$ is a generative model and it is not trained to discriminate between different classes; so, one should not expect to get better classification performance than a discriminatively trained model would give. Fortunately, these problems can be easily addressed by inverting $p(S|c)$ and assuming a uniform probability on classes:

$$p(c|r,S) \propto p(c|r)p(c|S)p(S). \quad (10)$$

The solution to semantic segmentation problem is then

$$c^* = \arg\max_c p(c|r,S) = \arg\max_c p(c|r)p(c|S) \quad (11)$$

where $p(S)$ is removed because it does not change the solution.

We train an image classifier based on the learned image statistics, $p(S;\Theta)$, as in the previous section. Then, $p(c|S)$ is estimated using this classifier by fitting a sigmoid function to raw classifier outputs.

We give per-class and average segmentation accuracies on the MSRC-21 dataset in Table 2. Note that the model with the prior (Eq. (9)) improves the average result by 8.6%. Although we believe that this improvement is sufficient to demonstrate the use of our proposed image prior model, we also include in Table 2 two results from state-of-the-art methods for comparison. With the prior model, our result is better than that of [31] despite the simplicity of our method. Both [31], [23] use sophisticated and costly random field models with many more features extracted per image than we do.

3.3 Object Detection

In this section, we demonstrate how our model helps improve the performance of an off-the-shelf object detector (we use [12]) on the PASCAL VOC 2007 dataset.

From an abstract point of view, an object detector is trained to compute $p(o|bb)$, that is the probability of o, an instance from a certain object class is observed, given a subimage bb, or some measurements extracted from this subimage. As we did in the previous section, when we use the image in addition to the subimage, we have: $p(o|bb,S) \propto p(o|bb)p(S|o)$.

To compute $p(o|bb)$, we take the confidence scores output by the object detector and simply scale them appropriately. For the image prior term, $p(S|o)$, we train an image classifier using the image statistics, as we did in the previous experiment. In our preliminary experiments, we noticed that the overall performance benefits from non-linearly scaling $p(S|o)$. In particular, we used an exponential scaling: $p(S|o)^\alpha, \quad \alpha > 0$. In the overall decision, the contribution of the image prior term is calibrated by α whose value is chosen by cross-validation on the `trainval` set.

We present the object detection performance in Table 3. The first row gives the detection results for the object detector alone, i.e. $p(o|bb)$. The second row shows the performance of the detector with the image prior, i.e. $p(o|bb,S)$. The use of the prior increases the detection performance by 3.3% on average over 20 classes. This is a larger improvement than the context-rescoring improvement proposed in [12]. Our object detection results are comparable with the state of the art results [12], [32] given in the last two rows of Table 3.

4 Discussion

We presented a set of statistics based on low-level segmentation of natural images. Based on these statistics, we were able to confirm that dominant orientations in natural images are horizontal and vertical. We also provided new findings such as that the number of regions versus photometric scale follows an exponential distribution, and that there are more regions in the lower halves of the images than there are in the upper halves. We proposed a MRF based model to learn the segmentation statistics and used this model in three high-level applications. One might ask why not directly optimize the segmentation for a given high-level task, i.e., semantic segmentation, as done in [15], [14], [7]. While doing so is a realistic and useful research direction to explore, it is not the main goal of this paper. Here we have aimed at obtaining a probabilistic model of photometric, geometric and topological structure of natural images, short of any semantics. Our approach is that tasks like semantic segmentation can follow our general segmentation as a follow up stage, and we have demonstrated that this separation is feasible. Doing so reduces the extra (combinatorial) complexity that would accrue from task specific segmentations - from having to pair tasks and segmentations for simultaneous optimization.

The main limitation of our model is that it is not a truly generative model for images, i.e. the model is not able to reconstruct the image. For this reason, for a class of problems where the output itself is a natural image, e.g. denoising, inpainting, etc., it is not trivial how to use our model. On the other hand, we believe that many high-level vision problems might benefit from the model proposed in this paper. We tried to demonstrate this in three simple applications.

Acknowledgments

The support of the Office of Naval Research (N00014-12-1-0259), National Science Foundation (IIS 11-44227) and Motorola Solutions (Motorola Communications Center grant 239, RPS #37) are gratefully acknowledged.

References

[1] N. Ahuja. A transform for multiscale image segmentation by integrated edge and region detection. *IEEE TPAMI*, 18(12):1211–1235, 1996.

[2] N. Ahuja and S. Todorovic. Connected Segmentation Tree: A joint representation of region layout and hierarchy. In *CVPR*, 2008.

[3] E. Akbas and N. Ahuja. From Ramp Discontinuities to Segmentation Tree. In *ACCV*, pp. 123–134, 2009.

[4] E. Akbas and N. Ahuja. Low-Level Image Segmentation Based Scene Classification. In *ICPR*, pp. 3623–3626, 2010.

[5] L. Alvarez, Y. Gousseau, and J.-M. Morel. Scales in natural images and a consequence on their bounded variation norm. In *Int'l Conf. on Scale-Space Theories in Computer Vision*, pp. 247–258, 1999.

[6] V. V. Appia and R. Narasimha. Low complexity orientation detection algorithm for real-time implementation. In *Proc. SPIE*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2299809, IEEE Transactions on Pattern Analysis and Machine Intelligence

	build	grass	tree	cow	sheep	sky	aero	water	face	car	bike	flower	sign	bird	book	chair	road	cat	dog	body	boat	**avg**
no prior	64.3	85.6	72.3	53.7	51.0	69.8	46.0	65.8	68.6	47.7	70.0	68.6	62.6	25.4	64.9	16.0	67.2	44.3	31.8	38.9	7.3	53.4
with prior	73.8	92.6	83.5	61.0	59.5	79.4	55.0	74.5	75.4	62.3	74.6	81.4	70.9	35.8	71.8	25.0	74.5	51.6	39.1	49.0	10.3	62.0
[31]	61.6	97.6	86.3	58.3	50.4	82.6	59.6	52.9	73.5	62.5	74.5	62.8	35.1	19.4	91.9	15.4	86.0	53.6	19.2	62.1	6.6	57.7
[23]	63.0	93.0	88.0	84.0	65.0	89.0	69.0	78.0	74.0	81.0	84.0	80.0	51.0	55.0	84.0	80.0	69.0	47.0	59.0	71.0	24.0	71.0

TABLE 2

Results and comparison on the MSRC-21 dataset.

	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**avg**
no prior	28.9	59.5	10.0	15.2	25.5	49.6	57.9	19.3	22.4	25.2	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6	32.3
with prior	33.4	61.6	13.4	19.2	28.6	52.3	60.2	23.2	26.8	28.2	27.4	13.8	60.2	51.8	44.2	16.6	22.1	36.6	47.7	44.7	35.6
[12]	31.2	61.5	11.9	17.4	27.0	49.1	59.6	23.1	23.0	26.3	24.9	12.9	60.1	51.0	43.2	13.4	18.8	36.2	49.1	43.0	34.1
[32]	38.6	58.7	18.0	18.7	31.8	53.6	56.0	30.6	23.5	31.1	36.6	20.9	62.6	47.9	41.2	18.8	23.5	41.8	53.6	45.3	37.7

TABLE 3

Object detection results in average precision on the PASCAL VOC 2007 dataset (trainval/test split). The second row ("with prior") shows the improvement when our image prior is used.

7871, Real-Time Image and Video Processing, 2011.

[7] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE TPAMI*, 34(7):1312–1328, 2012.

[8] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[9] T. S. Cho, N. Joshi, C. L. Zitnick, S. B. Kang, R. Szeliski, and W. T. Freeman. A content-aware image prior. In *CVPR*, 2010.

[10] D. M. Coppola, H. R. Purves, A. N. McCoy, and D. Purves. The distribution of oriented contours in the real world. *Proceedings of the National Acedemy of Sciences (PNAS)*, 95(7):4002–4006, 1998.

[11] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc 2007) results, 2007.

[12] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010.

[13] B. Ghanem, E. Resendiz, and N. Ahuja. Segmentation-based Perceptual Image Quality Assessment (SPIQA). In *ICIP*, pp. 393–396, Oct. 2008.

[14] S. Ghosh and E. B. Sudderth. Nonparametric learning for layered segmentation of natural images. In *CVPR*, pp. 2272–2279, 2012.

[15] S. Ghosh, A. B. Ungureanu, E. B. Sudderth, and D. M. Blei. Spatial distance dependent chinese restaurant processes for image segmentation. In *NIPS*, pp. 1476–1484, 2011.

[16] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, pp. 237–244, 2009.

[17] N. Hees, C. K. I. Williams, and G. E. Hinton. Learning generative texture models with extended Fields-of-Experts. In *BMVC*, 2009.

[18] J. Huang and D. Mumford. Statistics of Natural Images and Models. In *CVPR*, 1999.

[19] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer Computational Imaging and Vision, Vol. 39.

[20] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999.

[21] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer-Verlag, third edition, 2009.

[22] D. R. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pp. 416–425, 2001.

[23] D. Munoz, J. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *ECCV*, pp. 57–70, 2010.

[24] F. Perronnin and C. Dance. Fisher Kernels on Visual Vocabularies for Image Categorization. In *CVPR*, pp. 1–8. IEEE, 2007.

[25] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV (4)*, pp. 143–156, 2010.

[26] X. Ren and J. Malik. A probabilistic multi-scale model for contour completion based on image statistics. In *ECCV*, pp. 312–327, 2002.

[27] S. Roth and M. J. Black. Fields of Experts: A Framework for Learning Image Priors. In *CVPR*, pp. 860–867, 2005.

[28] S. Roth and M. J. Black. Fields of Experts. *IJCV*, 82(2):205–229, Jan. 2009.

[29] D. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548, 1994.

[30] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[31] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *IJCV*, 81(1):2–23, Dec. 2009.

[32] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, pp. 1585–1592, 2011.

[33] A. Srivastava, A. Lee, E. Simoncelli, and S.-C. Zhu. On Advances in Statistical Modeling of Natural Images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33–33, 2003.

[34] S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *IEEE TPAMI*, 30(12):2158–74, Dec. 2008.

[35] A. Torralba. Contextual Priming for Object Detection. *IJCV*, 53(2):169–191–191, 2003.

[36] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, Aug. 2003.

[37] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7), July 2009.

[38] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE TPAMI*, 32(7):1271–1283, 2010.

[39] L. Wang, X. Liu, L. Xia, G. Xu, and A. Bruckstein. Image orientation detection with integrated human perception cues (or which way is up). In *ICIP*, volume 2, pp. II – 539–42 vol.3, 2003.

[40] Y. Weiss and W. T. Freeman. What makes a good model of natural images? In *CVPR*, pp. 1–8, 2007.

[41] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. In *ICCV*, pp. 436–443, 2009.

[42] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32.1–20, Jan. 2008.

[43] D. Zoran and Y. Weiss. Scale invariance and noise in natural images. In *CVPR*, pp. 2209–2216, Sept. 2009.

Emre Akbas is a postdoctoral researcher in the Vision and Image Understanding Laboratory at the University of California Santa Barbara, where he is working on computational models of eye movements in visual search. He received his Ph.D. in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2011 and his M.S. degree in computer science from the Middle East Technical University in 2006.

Narendra Ahuja is the Donald Biggar Willet Professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. He is a Fellow of the IEEE, American Association for Artificial Intelligence, International Association for Pattern Recognition, Association for Computing Machinery, American Association for the Advancement of Science, and International Society for Optical Engineering. Narendra is on the editorial boards of several journals. He was the Founding Director of the International Institute of Information Technology, Hyderabad where he continues to serve as Director International. Narendra received his Ph.D. from the University of Maryland in 1979.