Labeling Spain With Stanford

Yingbo Zhou, *Student Member, IEEE*, Ifeoma Nwogu, *Member, IEEE*, and Venu Govindaraju, *Fellow, IEEE*

***Abstract*— We present an end-to-end framework for outdoor scene region decomposition, learned on a small set of randomly selected images that generalizes well to multiple data sets containing images from around the world. We discuss the different aspects of the framework especially a generalized variational inference method with better approximations to the true marginals of a graphical model. Experimentally, we explain why the framework is robust and performs competitively on many diverse scene data sets, including several unseen scene types. We have obtained high pixel-level accuracies (≈80%) in three of the four data sets, which include a benchmark data set known as the Stanford background data set. Our model obtained over 70% accuracy on the fourth data set, which contained a number of indoor and close-up images that are significantly different from our training examples.**

***Index Terms*— Scene understanding, semantic labeling, generalized mean field, generalization, low- and mid-level image cues.**

I. Introduction

MUCH of computer vision involves recognizing patterns in images and videos, one of the main challenges of pattern recognition algorithms is their ability to perform well when presented with new, not-previously-seen data. Performance results for different computer vision tasks such as event detection, scene recognition, object detection (with the exception of face detection) are typically presented in the context of a single dataset, and it is not always clear how such systems (both the algorithm and the model generated) will perform in new environments on the same problem. Hence, a central concept in both machine learning and computer vision is *generalization*: how to generalize beyond the examples provided during training to new examples presented during testing?

In this paper we approach the concept of generalization in the context of scene parsing (segmentation and annotation). We develop a stable and robust scene parsing algorithm trained on *only 572 images* and extend the learned model to test several significantly larger datasets including the Spain dataset.[1] The level of accuracy we have obtained in parsing the novel scenes indicates a robust generalization of our algorithm and suitable for modeling new environments.

The main contributions of this work are as follows: (i) we demonstrated the use of a cluster-based inference algorithm for scene parsing tasks in computer vision. The algorithm based on mean-field approximation is extended in a manner analogous to how belief propagation (BP) extends to generalized BP (GBP). We apply this inference method to image labeling and show experimentally how it improves our final results; (ii) we presented an *end-to-end framework* involving mid-level features, classifier and the usage of a region-optimal inference method, which allows us to generalize well across datasets; (iii) we introduced generalization benchmark results on three public datasets that can serve as a new measure of the generalizability of scene parsing methods. It is important to note that we are not stating that other methods do not generalize well, instead, going forward we encourage researchers in this area to perform similar benchmark tests on these diverse datasets for evaluating their generalization performance.

Fig. 1. Results of testing the scene decomposition task on different datasets with a model learned on 572 randomly selected images from Stanford BG dataset. Top: shadowed image from the Spain dataset; bottom: complex scene from the rest of the world dataset; Details of these and other datasets are given in Section IV. (images best viewed in color).

A. Related Work

In many high-level scene understanding tasks, it is important to specifically extract the regions present in a scene, in order to reason more accurately about the high-level concepts that exist in that scene. Hence, having a readily generalizable method of reasoning about scene regions and geometries, largely independent of training data, is an advancement in scene understanding. It is important to note that we distinguish between object recognition/categorization and scene decomposition problems. In object categorization, the scenes of interest are typically object-focused as observed from the benchmark object categorization datasets such as Pascal-VOC-2010 [1] and MSRC version 2 [2]. The goal here is to detect

Manuscript received February 6, 2013; revised June 19, 2013 and September 18, 2013; accepted October 2, 2013. Date of publication October 17, 2013; date of current version October 28, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikolaos V. Boulgouris.

The authors are with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14214 USA (e-mail: yingbo.zhou@ieee.org; inwogu@buffalo.edu; govind@buffalo.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2013.2285603

[1] http://people.csail.mit.edu/torralba/benchmarks/

Fig. 2. Examples highlighting challenges faced in scene region decomposition (best viewed in color).

specific foreground classes, under the assumption of a generic background class whereas scene decomposition is the inverted problem where the goal is to detect specific background classes, under the assumption of a single generic foreground class. Although it seems to believe that scene decomposition is a trivial problem when compared to object categorization (given that the relative smaller number of categories), there are challenges in accurate scene decomposition that include (i) the presence of one generic foreground class whose members tend to overlap significantly with other classes (for example, in Figure 2 the man's shirt shares similar cues as the background foliage). If this was a face or person detection problem, the existence of strong shape specific cues and priors would alleviate the task. However, this is a common problem in scene decomposition; (ii) also, unlike the typical object categorization problem, in scene region decomposition, different regions can share similar material and geometry cues and can tend to merge into each other (see Figure 2) make the region delineation challenging. We therefore limit our scope of comparisons to other works specifically in scene region decomposition.

One of the earlier works in scene region decomposition by Konishi and Yuille [3] used color and local texture filters to extract features from image pixels and classify them into one of six classes (edge, vegetation, air, road, building, and other). They constructed a probabilistic model using the empirical joint probability distributions of texture filter responses at multiple scales as well as prior knowledge about the typical number of each class per image. Finally, they applied Bayesian classification to assign a label to each pixel. The algorithm scored $>90\%$ for three of the six classes in their two datasets. The number of training and testing images from each of the two datasets was 50. We present a generalizable framework which is *trained on 572 images but generalizing to over 3,000 images from multiple sources*.

More recently, Gould *et al.* [4] presented a scene decomposition framework where they first partitioned their training images into multiple segmentations and the pre-computed segmentations were used to make a dictionary Ω of proposal moves for optimizing an energy function. Their energy minimization approach was defined in terms of multiple potentials: (i) the pixel-to-region association; (ii) the region semantic class; (iii) the region geometry; (iv) the region appearance; and (v) the location of the horizon. A two-stage variation of the iterated conditional modes (ICM) inference method was used to optimize the pixel-to-region assignment using dictionary Ω. This method could be somewhat viewed as a dynamically shifting conditional random field (CRF) where the the graphical model constantly changes its internal structure based on some external global energy criteria. Like the standard CRF, there are no explicit latent variables being estimated and the energy minimization considers both individual and pairwise potentials for inference. Along similar lines, Kumar and Koller [5] use the model of Gould *et al.* to compute unary and binary potentials. Unlike the previous method, a current iteration of the problem would merge and intersect with segments from the dictionary of regions to form putative regions. A tight linear program relaxation of an integer program was used to solve the energy minimization problem for this region-based model.

Lately, Socher *et al.* [6] have used a recursive neural network (RNN) with max-margin structure prediction to recursively identify the units of an image. They learned a pairwise score between adjacent segments and segments where the highest affinity scores were merged so that the underlying graphical structure was reconfigured to reflect the new "super-segment." By recursively pairing segments, a tree structure is implicitly defined over an image, where the root of the tree is the entire image. Each node in the tree has a feature representation associated with it. The class labels of the tree nodes are estimated by first defining a softmax class prediction for the nodes and then optimizing an error function across the entire neural network. Unlike the previous techniques, in our framework the underlying graphical structure does not change dynamically during inference, rather each node changes its label value during inference. Other interesting approaches preforming scene decomposition are described in [7]–[9].

II. Our Proposed Framework

Although there have been a few approaches to the scene region decomposition problem, a framework common to the more recent approaches includes (i) a segment (or superpixel) level image representation and the extraction of low-level cues especially color and texture from such segments; (ii) an underlying graphical structure over which class-specific probabilistic and/or energy models are learned; and (iii) a technique for assigning the learned class labels to every node in the graph which is then inherited by every pixel in the scene image. We therefore present our approach in the context of the common framework.

A. Image Representation and Low-Level Cues

Superpixels have become the representation-of-choice in many computer vision algorithms, especially for image labeling tasks [12]. The more recent scene decomposition methods described in Section I-A use superpixels either to generate the dictionary of proposal moves or as inputs to the training model. Since many superpixel generation algorithms were originally designed as full image segmentation algorithms, in complex scenes, they tend to yield superpixels which are highly irregular in shape and size. These then require heuristic-driven normalizations during processing. The Turbopixels algorithm [10] creates highly regular in shape and similarly sized regions.

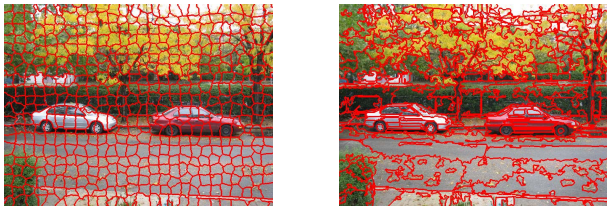

Fig. 3. Left: superpixels generated by Turbopixel [10]; right: segments generated by the local variation algorithm [11].

They are fast to compute and strongly respect local image boundaries - fulfilling much of our superpixel requirements.[2] Because a large part of our focus in our framework is on the underlying graphical structure, where the uniformity and regularity of the nodes is important, we opted to use the Turbopixels algorithm. Figure 3 shows a qualitative comparison of the algorithm with another widely used method [11] on a textured image.

We computed a 68-dimensional feature vector for each superpixel, consisting of color, texture, location and shape cues, which are a subset of the surface cues presented by Hoiem *et al.* in [12]. The color features are mean values from RGB, HSV, YCbCr and L*a*b* channels (12 dimensions); normalized histogram from hue (5 dimensions), saturation (3 dimensions), Cb (5 dimensions) and Cr (5 dimensions). The texture features are obtained from the mean absolute responses and histogram of maximum responses from 15 Leung-Malik filters [13] (*i.e.* total 30 dimensions). Location features are the mean horizontal and veritical locations, 10% and 90% x, y locations (6 dimensions). The shape features are ratios between the 10% and 90% x and y locations, along with the normalized area of the superpixel (2 dimensions).

We also computed a 37-dimensional pairwise feature for each pair of adjacent superpixels, consisting of the difference of the features of the two adjacent superpixels. All the differences were measured in terms of absolute values unless otherwise mentioned. Specifically, the features composed of the differences of the mean color values (12 dimensions); the Jenson-Shannon divergence between the normalized histograms (5 dimensions[3];) the difference between the mean texture responses (15 dimension); the difference between mean x, y locations (2 dimensions); the normalized area ratio between two superpixels (1 dimension); the ratio between the boundary length and the perimeter of the smaller superpixel (1 dimension); and the ratio between boundary length and endpoint distance (1 dimension).

B. Classifier and Mid-Level Cues

In addition to the low-level cues, empirically, we observed that although the horizon line was a very strong cue for labeling, it was also very unstable; *i.e.* when the horizon value shifted by a few pixels from the true value, it had a large negative impact on the resulting class labels. We therefore developed a geometry-based mid-level cue which we refer to as the *range-of-the-horizon* (as illustrated in Fig. 4). The range-of-the-horizon cue consists of the relative location of each superpixel to the two horizontal lines, one going through the lowest 10% superpixel in the sky plane, and the other going through the highest 10% superpixel in the support plane. So to obtain the range-of-the-horizon, we initially use *the same end-to-end framework* to parse an image into the geometry classes - sky, vertical, support (corresponding to the 3 main classes defined by Hoiem *et al.* [12]) and then compute the range lines.

Fig. 4. The two red lines in the image define the range-of-the-horizon cue.

We now combine this mid-level cue with the low-level cues from Section II-A so that the combined cues result in a 70-dimensional feature vector which are trained using boosted decision trees. To train the 8 semantic region labels, we trained 8 sets of boosted decision trees, in a one-versus-all fashion.

The pairwise classifier estimates the likelihood that two superpixels have the same label. This is a two-class problem where adjacent superpixels either have the same label or not. We obtain the training label from groundtruth and use the difference-based, pairwise feature that we computed in section II-A. Boosted decision trees are also used for classification here. The prediction from this pairwise classifier is used to generate the affinity matrix A.

The superpixels define an underlying undirected graph $G(V, E)$ whose edges are formed by the adjacency structure at each node $v_i \in V : \{i = 1, \cdots n\}$; n is the number of nodes in the graph. By imposing an exponential family distribution, we can apply an approximate mean-field-based inference which has been proven to converge to a globally consistent set of marginals and yields a lower bound on the likelihood [14]. Details of the inference algorithm are provided in Section III.

III. Regional-Optimal Graph Inference

A. Overview of Mean Field (MF) Approximation

Although the inference problem is tractable for graphical models with small treewidths, the general inference problem on graphs is NP-hard. For many graphical models of interest, especially in computer vision, the treewidth is too large to allow efficient exact inference, thus resulting in less accurate approximation methods. In this work, we use the variational approach to probabilistic inference. Because exact inference is infeasible in a problem such as this, we can convert our inference problem to an optimization problem by approximating the

[2]Turbopixel implementation can be found at http://www.cs.toronto.edu/~babalex/turbopixels_supplementary.tar.gz

[3]Corresponds to hue, saturation, Cb, Cr and texture histograms respectively

function to be optimized, and then solving the relaxed optimization problem. If $p(\mathbf{x}|\pi)$ is a probability distribution that factors according to a graph G, we can define an optimization problem that exploits the structure of G, so that the solution to the optimization problem results in approximations to the marginal probabilities. Such variational inference methods approximate $p(\mathbf{x}|\pi)$ with tractable distributions $q(\mathbf{x}|\alpha)$, where α are the set of free parameters. When $q(\mathbf{x}|\alpha)$ is a completely factorizable or tractable distribution, the class of methods is referred to as "mean field methods". A tractable family will correspond to a subgraph H of G.

When the subgraph has no edges, the approximating posterior distribution q can be completely factorized as:

$$q(\mathbf{X}) = \prod_i q_i(x_i). \tag{1}$$

B. Generalized Mean Field (GMF)

The relationship between generalized mean field (GMF) and the naive mean-field method can be viewed as being analogous to the relationship between generalized belief propagation (GBP) and ordinary belief propagation (BP).

GBP approximations [15] are performed on hypergraphs, defined over regular graphs via overlapping clusters of variables. The choice of clusters are often determined by the clique structure of the underlying graph, and go a long way in determining the performance of the GBP algorithm. Typically GBP algorithms require cluster-factorizability which is not always satisfiable for general distributions.

In a similar fashion, Xing *et al.* [14] introduced the class of GMF algorithms using *non-overlapping* clusters of cliques. Given a disjoint clustering of variables, $\mathcal{C} = \{C_1, C_2 \ldots C_I\}$, where C_i is the set of nodes in the ith cluster, and C_i need not form a clique; GMF defines a subgraph consisting of tractable connected components of clusters of nodes. Similar to the naive MF, the GMF approximation to the joint posterior can be expressed in as a product of tractable cluster marginals:

$$q(\mathbf{X}) = \prod q_i(\mathbf{X}_{C_i}) \tag{2}$$

By imposing an exponential distribution on the original graph:

$$p(\mathbf{X}|\theta) = \frac{1}{Z}\exp\{\sum_{\alpha\in\mathcal{A}} \theta_\alpha \phi_\alpha(\mathbf{X}_{D_\alpha})\} \tag{3}$$

$\mathcal{D} = \{D_\alpha|\alpha \in \mathcal{A}\}$ is the set of cliques of G indexed by the set $\mathcal{A}$; $\phi = \{\phi_\alpha|\alpha \in \mathcal{A}\}$ is the set of clique potentials; $\theta = \{\theta_\alpha|\alpha \in \mathcal{A}\}$ is the set of parameters associated with the node potential functions ϕ; and Z is the normalization constant.

Hence, for a disjoint clustering of variables, $\mathcal{C}$, the true cluster conditional for a cluster $C_i \in \mathcal{C}$ can be written as:

$$p(\mathbf{X}_{C_i}|\mathbf{X}_{MB_i} = \mathbf{x}_{MB_i}) \propto \exp\{\sum_{D_\alpha\subseteq C_i} \theta_\alpha\phi_\alpha(\mathbf{X}_{D_\alpha}) + \sum_{D_\beta\subseteq B_i} \theta_\beta\phi_\beta(\mathbf{X}_{D_\beta\cap C_i}, \mathbf{x}_{D_\beta\cap MB_i})\} \tag{4}$$

MB_i is the Markov blanket of C_i; B_i is the set of cliques intersecting with C_i but not contained in C_i (see Figure 5);

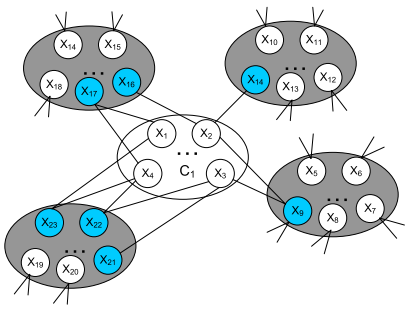

Fig. 5. The Markov blanket MB_1 of cluster C_1 are the blue shaded nodes while the gray blobs are cliques intersecting with C_i but not contained in it.

and lowercase $\mathbf{x}$ is a specific assignment to $\mathbf{X}$.

For a given clique D_β, let I_β be the indices of clusters intersecting with the clique D_β so that $I_{\beta_i} = I_\beta \backslash i$. By defining a term referred to as the *peripheral marginal potential* of cluster C_i given by $\phi'_\beta(\mathbf{X}_{D_\beta\cap C_i}, q_{I_{\beta_i}})$, the GMF approximation of the cluster marginal $q_i(\mathbf{X}_{C_i})$ is isomorphic to the true cluster conditional of Equation 3 [16].

This isomorphism allows for an asynchronous iteration procedure looping over each cluster, calculating its peripheral marginal potentials using the current cluster marginals of its own Markov blanket clusters, and then updating its own cluster marginal. The advantage of such a technique is that more efficient inference (even exact inference) can be done inside each cluster (region), so that the final approximation would be more accurate.

C. Using Graph Partitioning for Variable Clustering

Similar to generalized belief propagation, the quality of the generalized mean field approximation depends critically on the choice of variable clustering of the underlying graph G. Unlike generalized belief propagation which involves the generation of hypergraphs of overlapping clusters from G, GMF involves non-overlapping clusters, thus making the application of graph partitioning methods very attractive. For clustering variables it will be desirable to break up cliques with small weights, hence we consider graph partitioning based on minimum costs.

If we now consider the graph generated by our superpixel organization, $G = (V, E)$ with a nonnegative edge function $a : E \leftarrow [0, \inf)$, a k-partition of V is a collection $P = \{V_1, V_2, \ldots V_k\}$ of k disjoint subsets of V, whose union equals to V. The symmetric matrix $A = \{a_{ij}\}$ is the affinity matrix where $a_{ij} = 0$ when there is no edge, and $a_{ii} = 0, \forall i$.

If we define an $n \times k$ matrix $\Pi = \pi_{vj}$ as the *k-partition matrix* where $\pi_{vj} \in \{0, 1\}, \forall v, j$, then Π is an orthogonal matrix and $\|\Pi\|_F = \sqrt{n}$. The notation $\|\cdot\|_F$ is the Frobenius norm, *i.e.* $\|\Pi\|_F = \sqrt{\text{trace}\Pi^T\Pi}$.

Hespanha [17] showed that an $n\times k$ matrix Π is a k-partition matrix if and only if each row of Π is a vector of the canonical basis of $\mathbb{R}^k$. Hence a k-partition matrix Π is completely specified by an n-vector whose vth entry contains the index within the vth row of Π of the entry equal to one. This vector is called the *partition vector* associated with Π. There is a one-to-one correspondence between the set of k-partitions of $V := 1, 2, \ldots, n$ and the set of k-partition matrices Π, so that the graph partitioning problem of finding a

minimum cost k-partition of G (no partition should have more than l nodes and $l = \lceil n/k \rceil$ for perfect balance, although in practice a small error of imbalance ϵ is tolerated) is cast as.

$$\begin{aligned} &\text{maximize trace}(\Pi^T A \Pi) \\ &\text{subject to } \pi_{vj} \in 0, 1, \forall v, j \\ &\Pi \text{ orthogonal}, \quad \mathbf{1}_n \Pi \mathbf{1}_k = n, \quad \mathbf{1}_n \Pi e_i < l, \end{aligned} \tag{5}$$

where $\mathbf{1}_n$ is an n-vector with all entries equal to one and e_i is the canonical basis of $\mathbb{R}^k$. The above maximization problem can be very efficiently solved using spectral clustering methods, details of which are provided in [17].

D. Our Implementation of GMF and Spectral Graph Partitioning

In order to implement generalized mean field (GMF) inference, we first generate an undirected graph from the superpixels (by using the Turbopixel algorithm [10]). The selection of l in equation 5 is critical for the success of the GMF inference. Too large (or small) values of l decreases the efficiency of the algorithm since it reduces the GMF back to naive MF. We select parameter l using a heuristic so that the resulting combined nodes (superpixels) can cover a small region that represent one class in the image. For example, let s be the size of the superpixel and α be the number of pixels that can roughly cover an area of the same label, then we set $l \leq \alpha/s$. In practice, the selection of l is not very sensitive, we found similar performance of different choices of l provided that the previous condition is met. For the nodes $v \in V$, the output of the boosted decision tree classifier was selected as the single node potential θ_i. The pairwise potentials were recorded in the affinity matrix A (see section II-B). The inputs to the GMF are therefore (i) the learned classification model $p(\mathbf{X}, \mathbf{Y})$ where $\mathbf{X}$ are single node potentials and $\mathbf{Y}$ are the class labels, (ii) the node pairwise relationships in the form of the affinity matrix A; and (iii) the set of l-bound partitions (or clusters) $\mathcal{C}$ obtained from the spectral graph partitioning method. The resulting output is the approximation $q(\mathbf{Y})$.

In the next section, we present the results we obtained using this improved inference approximation, on several diverse benchmark datasets obtained from different parts of the world.

IV. Experiments and Results

To show the extent to which our proposed framework generalizes, we perform the scene decomposition task on several datasets including the Stanford BG dataset [4] which has been tested extensively using various methods. **We ran the scene parsing tests on all the other datasets using only the model learned from the single best performing fold (572 training images) of the Stanford BG test**. Hence, we used the model learned from 572 training images to successfully label over 3,000 other diverse scenes from different parts of the world across. To illustrate the effectiveness of the GMF method, we comparatively tested it against both its naive counterpart, the single node potential baseline and the region-based energy method [4]. In the single node potential case, no pairwise information is used in the inference. A summary of the results obtained from all the datasets is given in Table I.[4]

TABLE I

Results (Best-Fold) Comparing Generalized Mean Field (GMF) With Other Inference Methods, Where the Training Data Is from Stanford BG and Test Data Is from the Datasets in the Leftmost Column

Dataset	Accuracy			
	Single node potential	MF	GMF	Region-based energy [4]
Rest of the world:	66.10%	68.36%	70.97%	67.28%
Spain:	73.48%	76.53%	79.34%	70.56%
make3D:	75.04%	78.73%	81.27%	77.03%
Stanford BG:	75.62%	79.05%	81.03%	78.59%

TABLE II

Confusion Matrix from 5-Fold Cross Validation on Dataset Classification

	Stanford BG	Spain	ROTW	Make3D
Stanford BG	**81.32%**	1.51%	14.15%	3.02%
Spain	4.76%	**13.57%**	62.52%	19.16%
ROTW	2.52%	2.96%	**83.87%**	10.65%
Make3D	4.07%	7.88%	57.88%	**30.18%**

A. Datasets

In order to illustrate the effectiveness of the proposed framework, we tested on the following four datasets: the Stanford background dataset [4], make3D dataset [18], [19], Spain dataset, and rest of the world dataset (ROTW) from LabelMe [20]. To further illustrate the distinctness of the different datasets, we carried out an experiment similar to the one mentioned in [21]. For all four datasets, we extracted GIST features [22] and then trained a linear support vector machine to classify each dataset. We performed 5-fold cross validation and obtained 59.85% classification accuracy and the confusion matrix of the four datasets are shown in table II. As can be observed from the confusion matrix, the overlap of the Stanford dataset with other datasets is relatively small.

B. Overview of Results

1) The Stanford Dataset: The Stanford BG dataset consists of 715 fully annotated outdoor scene images, where each image contains at least one foreground object and has the horizon positioned within the image [4]. The standard test involves a 5-fold cross validation, where the dataset is randomly split into 572 training images and 143 test images for each fold. Average pixel level accuracies over the 5 folds are reported. Our framework outperforms all the state-of-the-art techniques on this dataset as shown in Table III.[5] Examples of labeled images as well as the confusion matrix are shown in Figure 6 and 7, respectively.

[4]These results were obtained by taking the best model in the 5-fold cross validation and applying it to the rest of the datasets. Note: while some tables report the best performance, others report average performances, but this is indicated in the caption.

[5]Kumar and Koller [5] are not included in the table as they did not report results for a 5-fold cross validation test.

TABLE III
RESULTS (AVERAGE ACCURACY) COMPARING GENERALIZED MEAN FIELD (GMF) WITH OTHER INFERENCE METHODS REPORTED IN THE LITERATURE, WHERE THE TRAINING DATA IS FROM STANFORD BG AND TEST DATA IS FROM STANFORD BG

Method	Pixel Accuracy
Pixel CRF, [4]	74.3%
Region-based energy [4]	76.4%
Local labeling [23]	76.9%
Superpixel MRF [23]	77.5%
Simultaneous MRF [23]	77.5%
RNN [6]	78.1%
Our framework (MF)	77.3%
Our framework (GMF)	**80.2**%

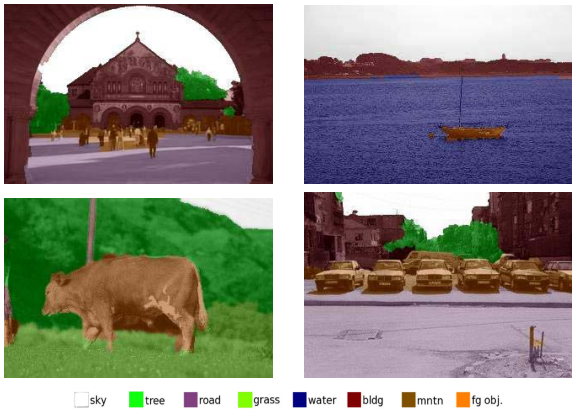

Fig. 6. Examples of labeled images from Stanford dataset. (images best viewed in color).

2) The Spain Dataset: The Spain dataset was originally collected as the training data for the benchmark dataset used in recognizing and segmenting as many object categories as possible. The dataset is a compilation of outdoor pictures taken in different cities of Spain. It was annotated using the LabelMe [20] annotation tool. We unified our labels with those from LabelMe by renaming the major classes object, person and cars to our generic foreground class and we also combined Sidewalk and Road into the Road class. We renamed a total of 141 classes and mapped them into the 8 semantic classes as Stanford BG dataset. The label mapping is provided in the appendix.

The dataset contained 2,920 images; after processing the dataset and getting rid of unidentifiable images, we were left with 2,301 images for testing. Not all images were fully annotated for this dataset. It contains more than 1,000 fully annotated images and about 2,000 partially annotated images.

For processing, each image in the Spain dataset was resized so that the longer side of the image occupies 320 pixels, where the aspect ratio of the image remains unchanged.The turbopixel algorithm was applied to yield roughly 500 superpixels per image and the chosen robust features were extracted from the superpixels. Two classification models (geometry model and semantic region model) learned from training on the 572 images randomly selected from the Stanford BG (that accounted for the highest accuracy among the 5 folds) were applied to compute the mid-level cues and to initially classify the superpixels. The final labels were assigned during inference and every pixel inherited the label of its parent superpixel. Pixel-level accuracy was computed by comparing the assigned label after inference to the label provided by the LabelMe annotations. The accuracy we obtained on the Spain dataset was **79.34%**. Examples of labeled images and confusion matrix are shown in Figure 8 and 7, respectively.

3) The "Rest of the World" (ROTW) Dataset: The ROTW dataset was originally collected as the testing data counterpart of the Spain dataset described in Section IV-B.2. It consists of images taken from the rest of the world other than Spain so that the dataset bias can be minimized. We use this intentionally designed diversity of the datasets to test the generalization power of our scene decomposition framework. ROTW was also annotated using LabelMe, and the dataset labels and images were unified in a similar manner as before and a total of 135 categoried were renamed. The label mapping of this dataset is also provided in appendix.

The dataset initially contained 1,133 images, some indoors, close up shots of objects, and others yet in scenes not previously seen by our training model. Unlike before, we did not get any unidentifiable images, using the same processing method as described for Spain. The accuracy we obtained on the ROTW dataset (all 1,133 images) was **70.97%**. Examples of labeled images are shown in Figure 9.

4) Semantically-Augmented Make3D Dataset (Make3D): The Make3D dataset [18], [19] contains images and depthmaps obtained from the Make3D project. The dataset is further extended in Liu *et al.* [24] for estimating depth in single images using predicted semantic labels. These semantic classes correspond 1-to-1 with the Stanford Background dataset. Make3D consisted of 534 fully annotated images. The pixel-level accuracy we obtained on this dataset was **81.27%**. Examples of labeled images are shown in Figure 10.

5) Additional Images from Multiple Datasets: In Figure 1 and 11 we show additional challenging images whose contents were very dissimilar to any of our training images, yet the resulting labels were still good.

V. DISCUSSION

The confusion matrices from Stanford BG and Spain dataset are quite similar (see Figure 7), and the classes that perform worse are water, mountain and the foreground objects. It is not suprising that "mountain" gets the lowest accuracy since in Stanford dataset there are only a few images that contain mountain. The performance of foreground objects is also as expected, since it contains a variety of classes which have significantly different features. In addition, the foreground objects tend to have shared features with background classes. The horizon line seems to be a very good cue for foreground objects, as we obseved in many images most foreground objects are near or below the horizons. This can also be observed from the confusion matrices – the classes that mixed with foreground objects are primarily road and building. Road almost always falls below the horizon and buildings typically intersect the horizon.

Both approaches that use pairwise potentials out performs the one that only utilizes single node potentials, which is

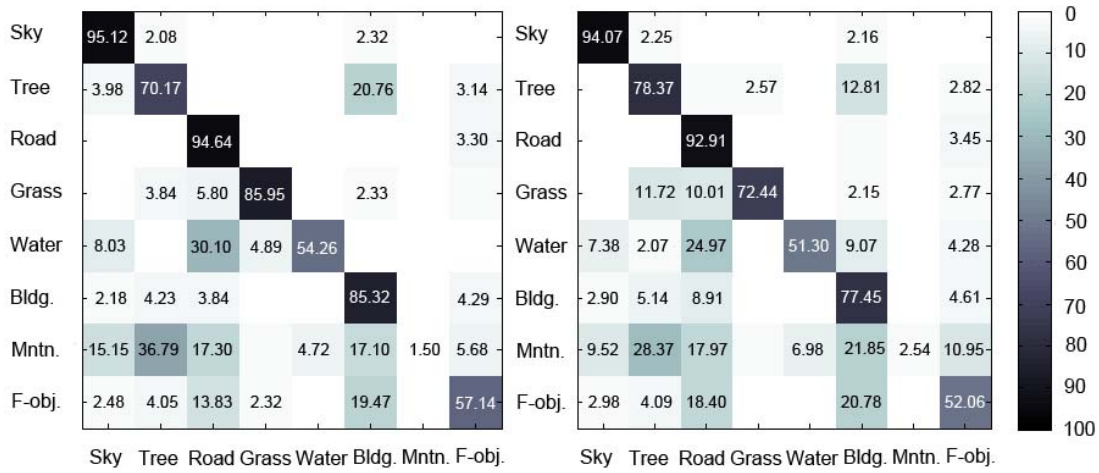

Fig. 7. Labeling Spain with Stanford: confusion matrices for the Stanford BG dataset (left) and the Spain dataset (right).

TABLE IV
ILLUSTRATION OF HOW ACCURACY OF LABELING VARIES WITH CHOICE OF FEATURES

	Overall Accuracy	Sky	Tree	Road	Grass	Water	Building	Mountain	Foreground Object
Location Only	56.03%	80.86%	12.28%	91.87%	18.14%	22.78%	71.54%	0.00%	15.47%
Color Only	62.70%	85.15%	48.77%	73.90%	78.82%	23.92%	69.47%	0.00%	27.55%
Texture Only	65.24%	76.78%	63.82%	83.81%	44.35%	5.99%	83.66%	0.00%	16.24%
No Color	73.28%	89.48%	65.62%	92.57%	44.24%	31.64%	81.18%	0.00%	47.52%
No Texture	75.78%	92.53%	53.43%	92.73%	82.40%	40.32%	83.25%	0.00%	49.93%
No Location	78.90%	93.23%	67.77%	93.44%	86.79%	47.79%	82.81%	0.00%	53.59%
No Range of Horizon	79.34%	93.81%	68.51%	93.70%	86.76%	49.66%	83.42%	0.00%	53.45%
All Features	81.03%	95.12%	70.17%	94.64%	85.95%	54.26%	85.32%	1.50%	57.14%

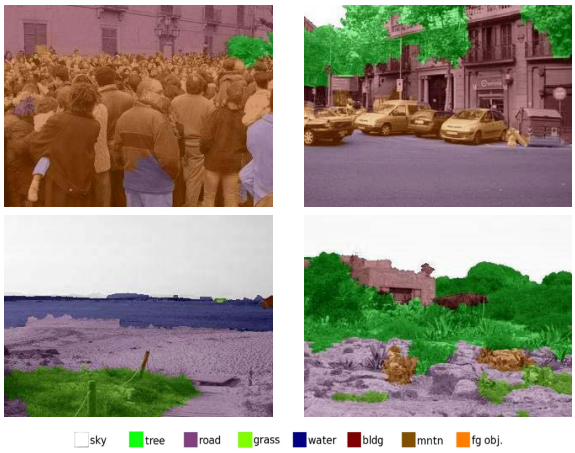

Fig. 8. Examples of labeled images from Spain dataset. (images best viewed in color).

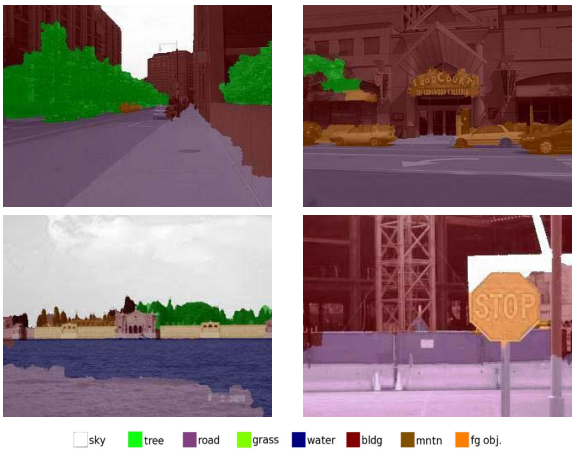

Fig. 9. Examples of labeled images from "rest of the world" dataset. (images best viewed in color).

as expected. However, merely applying inference on both single and pairwise potentials tends to give solutions that are over smoothed because of the strong indications of smoothness preservation in the pairwise term. By using our generalized mean field, we are able to utilize larger regions to do inference. So, on the one hand it enabled us to get smooth region labels as before inside the combined nodes, since they all have similar properties. On the other hand, it also allows us to capture discontinuity across large boundries, because the combined region is relatively large, which suppressed the pairwise term so that we do not get over smoothed. The region boundries from various datasets are quite clean (see examples in figure 6, 8 – 11), and the label is smooth inside large regions.

A. Analysis of Features

We experimentally evaluated how the choice of features affect the accuracy obtained from again labeling a single run of the Stanford BG dataset and present the results in Table IV. We observe that only the large homogenous regions such as the sky and grass (from afar) perform better with color only

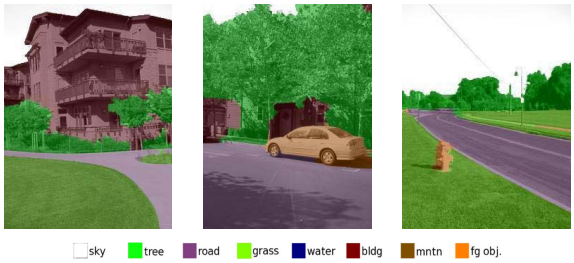

Fig. 10. Examples of labeled images from Make3D dataset. (images best viewed in color).

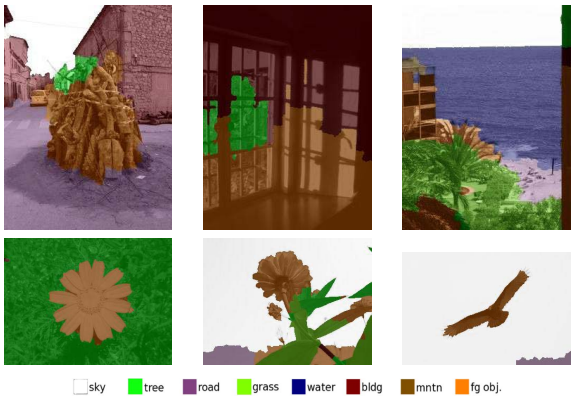

Fig. 11. Labeling examples from multiple datasets. (images best viewed in color).

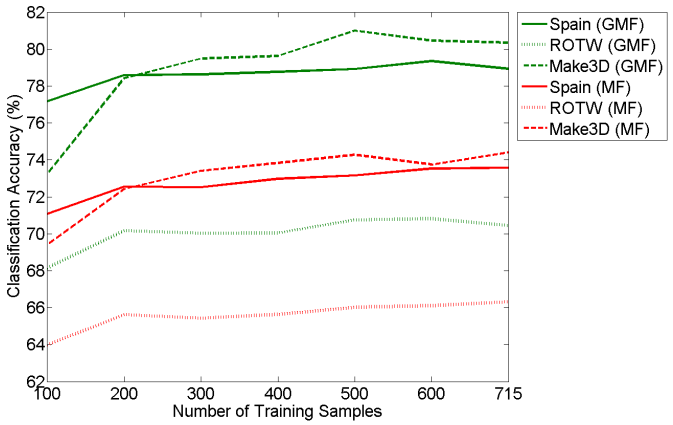

Fig. 12. Results (best-fold) of comparing generalized mean field (GMF) with naive mean field (MF), when varying the numbers of training examples. The training data is from Stanford BG and test data is from the three other datasets shown in the legend.

cue than with texture only cue, since in homogenous regions texture will not make much difference. Also, location specific regions such as sky, road, water and building perform quite well with location-based cues as the only set of features. As expected, removing either texture or color significantly impacts the performance of all the classes as well as the overall accuracy. As mentioned previously, the range of horizon features are efficient for foreground objects, and from the results we observe, it also improved the performance in several other classes. A possible reason for the good performance of this feature is that the relative locations among regions are stable

TABLE V

RESULTS (BEST-FOLD) OF COMPARING GENERALIZED MEAN FIELD (GMF) WITH THE REGION-BASED METHOD [4], WHERE THE TRAINING DATA IS FROM THE MAKE3D DATASET AND TEST DATA IS FROM THE DATASETS ON THE LEFTMOST COLUMN

Dataset	**Accuracy**	
	GMF	Region-based energy [4]
make3D:	86.25%	84.44%
Rest of the world:	62.05%	55.53%
Spain:	65.74%	54.63%
Stanford BG:	69.27%	57.77%

but the absolute locations might have high variations. With the combination of all the features, the best accuracy value is obtained and the framework outperforms all other reported results on the same dataset.

B. Analysis of Performance

We experimentally evaluated how the number of training examples influenced the performance of scene region decomposition by varying the number of training images selected from the Stanford BG dataset and running the same classification tests on the other datasets. We selected seven different training subsets from the Stanford BG dataset, where the first six subsets were randomly selected, with sizes varying uniformly from 100 to 600. The last subset simply included all 715 training images from the Stanford BG dataset. We evaluated our proposed model as well as a mean field baseline model for all the cases, and the results are illustrated in Figure 12. The performance increased much slower after 200 training examples, and stabilized after 400 training examples. It is clear from these results that the proposed method using GMF consistently outperformed its naive counterpart, for each of the datasets. Also interesting is the fact that the shape of the curve was quite similar for each dataset, although GMF was consistently higher.

To further test the robustness of the GMF method, we also evaluated the performance by training on make 3D and testing on the other datasets. We followed the same procedure as described in Section IV, *i.e.* applying 5-fold cross validation and using the single best model to test on the rest of the datasets. It is clear from the results (see Table V) that our method consistently performs better even across datasets.

VI. CONCLUSION

We have presented a robust scene parsing framework that generalizes well to diverse datasets of outdoor scenes. We attribute this robustness to our framework, the mid-level feature based on the location of the sky and support (or ground) planes and an improved inference method. Also, from a segmentation-only perspective, our choice of initial superpixels resulted in segmented regions that were true to their original boundaries. In addition, the regional based inference enabled us to get more accurate approximation for the probabilities of the underlying graphical model.

As expected, the identification of the generic foreground class whose members tended to overlap significantly with other classes, is challenging. The corresponding average accuracies are at <60% across the datasets.

We believe that a generalizable scene decomposition framework such as the one we have presented is a significant advance in scene understanding. However, to complete the loop, it will be important to incorporate the problems of object localization and categorization with scene decomposition, so that strongly identifiable scene regions can act as priors for the location and position of weakly identifiable objects and vice versa. Rather than attempting to perform these two separate tasks concurrently, it might prove more useful to perform them iteratively, where one set of cues can enhance the existence of the other. The horizon feature proved quite useful, hence, going forward, it will be interesting to incorporate it into the inference process in the CRF model.

APPENDIX

SEMANTIC LABEL MAPPING FROM LABELME TO STANFORD BG LABELS

TABLE VI
LABEL MAPPING FOR SPAIN DATASET

Stanford Semantic Labels	LabelMe Labels
Foreground	aerial, air conditioning, arcade, arrow, bag, bench, bicycle, billboard, bird, blind, boat, bottle, box, brand name, brushes, bus, camera, car, car az0deg, car az120deg, car az150deg, car az180deg, car az210deg, car az240deg, car az270deg, car az300deg, car az30deg, car az330deg, car az60deg, car az90deg, central reservation, chain, chair, clock, crane, decoration, dock, dog, duck, face, flag, head, headlight, hung out, knob, lamp, lantern, license plate, light, mailbox, mirror, moped, motorbike, motorcyclist, number, people, person, person az0deg, person az120deg, person az150deg, person az180deg, person az240deg, person az270deg, person az300deg, person az30deg, person az60deg, person az90deg, phone, pipe, plaque, plastic, pole, poster, pot, sculpture, sheep, shoe, sign, stand, stand occluded, step, streetlight, table, tail light, text, traffic light, trash, truck, umbrella, van, vase, wheel, windshield, wire, mast
Building	awning, balcony, balustrade, booth, building, bus stop, chimney, church, column, door, fence, gate, handrail, mill, pane, railing, railings, roof, sidewalk cafe, stair, tower, wall, window
Road	bridge, crosswalk, curb, floor, ground, path, road, sand, sidewalk, manhole
Water	sea, water, sping
Sky	cloud, sky
Tree	tree, leaf, plant
Mountain	mountain, rock
Grass	field, flower, grass

TABLE VII
LABEL MAPPING FOR REST OF THE WORLD DATASET

Stanford Semantic Labels	LabelMe Labels
Foreground	aerial, arm, bag, bench, bicycle, bike, billboard, bird, blind, boat, box, bus, butterfly, car az0deg, car az120deg, car az150deg, car az180deg, car az210deg, car az240deg, car az270deg, car az300deg, car az30deg, car az330deg, car az60deg, car az90deg, car_right, car_top_back, car_top_front, chair, car, clock, cone, duck, dummy, eye, face, firehydrant, flag, hat, head, headlight, ladder, leg, license plate, lion, mailbox, mannequin, mast, motorbike, newspaper stand, painting, parkingmeter, person, person az0deg, person az120deg, person az180deg, person az240deg, person az270deg, person az30deg, person az90deg, person occluded walking, picture, pigeon, pipe, pole, poster, pot, sculpture, sign, spare tire, squirrel, stand, streetlight, table, tail light, torso, traffic light, trash, umbrella, van, truck, wheel, windshield, windshield occluded, wire, zebra
Building	attic, awning, balcony, booth, building, chimney, column, door, entrance, fence, gate, roof, handrail, grille, grille occluded, pane, railing, railings, tower, wall, widnow, window, stair
Road	bridge, block, crosswalk, curb, ground, guard rail, path, road, sand, sidewalk
Water	lake, river, sea, water
Sky	cloud, sky
Tree	branch, leaf, plant, tree
Mountain	mountain, rock
Grass	field, flower, grass

REFERENCES

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2010). *The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results* [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html

[2] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, Jan. 2009.

[3] S. Konishi and A. L. Yuille, "Statistical cues for domain specific image segmentation with performance analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2000, pp. 1125–1132.

[4] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1–8.

[5] M. P. Kumar and D. Koller, "Efficiently selecting regions for scene understanding," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 3217–3224.

[6] R. Socher, C. C.-Y. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proc. 26th ICML*, 2011, pp. 129–136.

[7] L.-J. Li, R. Socher, and F.-F. Li, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 2036–2043.

[8] X. He, R. Zemel, and M. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proc. IEEE Conf. CVPR*, vol. 4. Jun./Jul. 2004, pp. 695–702.

[9] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 113–140, Jul. 2005.

[10] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "TurboPixels: Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2290–2297, Dec. 2009.
[11] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.
[12] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 151–172, Jan. 2007.
[13] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29–44, Jun. 2001.
[14] E. P. Xing, M. I. Jordan, and S. Russell, "A generalized mean field algorithm for variational inference in exponential families," in *Proc. 19th Conf. Uncertainty Artif. Intell.*, 2003, pp. 583–591.
[15] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.
[16] E. P. Xing, M. I. Jordan, and S. Russell, "Graph partition strategies for generalized mean field inference," in *Proc. 20th Conf. Uncertainty Artif. Intell.*, 2004, pp. 602–610.
[17] J. P. Hespanha, "An efficient MATLAB algorithm for graph partitioning," Dept. Elect. Comput. Eng., Univ. California, Santa Barbara, CA, USA, Tech. Rep., Oct. 2004 [Online]. Available: http://www.ece.ucsb.edu/ hespanha/published/tr-ell-gp.pdf
[18] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. Adv. NIPS*, 2005, pp. 1–8.
[19] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.
[20] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabeLME: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, May 2008.
[21] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 1521–1528.
[22] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[23] J. Tighe and S. Lazebnik, "SuperParsing: Scalable nonparametric image parsing with superpixels," in *Proc. ECCV*, Sep. 2010, pp. 352–365.
[24] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 1253–1260.

Yingbo Zhou received the B.E. degree in computer science from the Civil Aviation University of China, China, in 2006, and the M.S. degree in computer science (with distinction) from the Hong Kong Polytechnic University, Hong Kong, in 2010. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY, USA. His research interests include machine learning, computer vision, and artificial intelligence.

Ifeoma Nwogu received the B.S. degree from the University of Lagos, Lagos, Nigeria, the M.S. degree from the University of Pennsylvania, Philadelphia, PA, USA, and the Ph.D. degree from the University at Buffalo, State University of New York, Buffalo, NY, USA, as a U.S. National Science Foundation Graduate Fellow, in 2009, in electrical engineering, computer and information science and computer science and engineering, respectively. She joined the Computer Science Department, University of Rochester, Rochester, NY, USA, from 2009 to 2011, as an NSF Computing-Innovations Fellow and is currently a Research Scientist at the University at Buffalo. Her work is supported by the NSF, the Department of Defense and industry-based grants, and her research interests include image understanding, statistical learning methods, and artificial intelligence.

Venu Govindarajul is a SUNY Distinguished Professor of computer science and engineering with the University at Buffalo (UB), State University of New York, Buffalo, NY, USA. He received the B.Tech. (Hons.) degree from the Indian Institute of Technology (IIT), Kharagpur, India, and the Ph.D. degree at UB. He is the Founding Director with the Center for Unified Biometrics and Sensors. He is a recipient of the IEEE Technical Achievement Award, and is a fellow of AAAS, ACM, IAPR, and SPIE.