Scenes vs. Objects: a Comparative Study of Two Approaches to Context Based Recognition

Andrew Rabinovich
Google Inc.
amrabino@google.com

Serge Belongie
University of California, San Diego
sjb@cs.ucsd.edu

Abstract

Contextual models play a very important role in the task of object recognition. Over the years, two kinds of contextual models have emerged: models with contextual inference based on the statistical summary of the scene (we will refer to these as Scene Based Context models, or SBC), and models representing the context in terms of relationships among objects in the image (Object Based Context, or OBC). In designing object recognition systems, it is necessary to understand the theoretical and practical properties of such approaches. This work provides an analysis of these models and evaluates two of their representatives using the LabelMe dataset. We demonstrate a considerable margin of improvement using the OBC style approach.

1. Introduction

In the computer vision community, contextual models for object recognition were introduced in late 1980's and early 1990's [1, 6, 11], and were popularized by Oliva and Torralba in 2001 [7]. While they employ a variety of formulations, most of the approaches can be classified into two general categories: (i) models with contextual inference based on the statistical summary of the scene (we will refer to these as scene based context models, or SBC), and (ii) models representing the context in terms of relationships among objects in the image (object based context, or OBC).

The approach of [7], later termed Gist [12], was fundamental among the SBC models. Since then, variants of the SBC model were presented in [3, 5, 14, 15]. These recent works have shown that a statistical summary of the scene provides a complementary and effective source of information for contextual inference, which enables humans to quickly guide their attention to regions of interest in natural scenes.

SBC models of context, Gist-based approaches in particular, aim to capture the surrounding information around the object of interest. By incorporating the statistics of the clutter or background, context becomes a global feature of the object category. For example, refrigerators usually appear in a kitchen, thus the usual background of refrigerators is similar. Having learned such a global feature of an object category, one can infer a potential object label: if the background resembles a kitchen, then the patch of interest may be a refrigerator. However, many objects can have similar backgrounds, e.g., refrigerators, coffee makers, and stoves all belong in the kitchen. Alternatively, instances of a particular object (a face or a car), may have very different backgrounds depending on the environment they are in. Faces, for example, may appear outdoors or inside, at night or during the day. As illustrated in Figure 1(a,c), the background of an object may not always be indicative of the object itself.

Proceeding with the SBC model, after measuring the global features of the image, one first infers the scene context of the image, e.g., kitchen, and then with scene context in hand, the label of the object is inferred, e.g., refrigerator. Notice that if the scene context is inferred incorrectly, it becomes impossible to identify the object label accurately.

(a) (b) (c)

Figure 1. The structure of objects and their backgrounds (taken from [12]). In this illustration, each image has been created by averaging hundreds of images containing a particular object in the center (a face, keyboard and fire hydrant) at a fixed scale and pose. Before averaging, each image is translated and scaled so that the target object is in the center. The averages can reveal the regularities existing in the color/brightness patterns across all the images. However, this behavior is only visible for the keyboard in (b). In (a), the background of a face is approximately uniform, since faces appear in a variety of settings. Alternatively in (c), the background of a fire hydrant, may be identical to that of a bus stop or a street sign.

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

An alternative approach to Gist and other SBC models is to use a method based on the OBC model, variants of which were presented by [2, 8]. Rather than measuring global image statistics, inter-object constraints are imposed on potential object candidates in the image. With learned category interaction probabilities, either from training data or generic sources on the web, object labels are given to image regions, such that mutual co-occurrence and spatial constraints among all the object labels in the image are maximized. In OBC approaches, only the object category labels must be inferred given the context between categories and individual object appearance, without regard for scene context. To illustrate this further, turn to the example of an idealized OBC model in Figure 2, taken from [8].

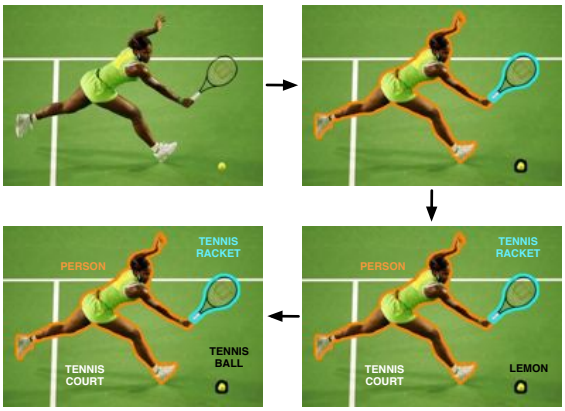

Figure 2. A possible idealized model for object recognition. An original image is segmented into objects; each object is categorized; and object labels are adjusted with respect to semantic context in the image. As a result, the label of the yellow blob changes from "Lemon" to "Tennis Ball".

In the scene of a tennis match, four objects are detected and categorized: "Tennis court", "Person", "Tennis Racket", and "Lemon". Using a categorization system without a contextual module, these labels would be final; however, in context, one of these labels is not satisfactory. Namely, the object labeled "Lemon", with an appearance very similar to a "Tennis Ball" is mislabeled due to the ambiguity in visual appearance. By modeling context with OBC constraints provided by an oracle, the label of the yellow blob changes to "Tennis Ball," as this label better satisfies the contextual conditions. While the above mentioned formulations of context appear rather different, it is clear that inclusion of context, in some form, in object recognition is a must. Thus, we are faced with a dilemma: which contextual model is more suitable in the framework of automated object recognition or categorization? Furthermore, which model is simpler, and finally, do the differences in the formulations matter? In the following sections, we pose both SBC and OBC models in a manner most suitable for a comparison and an evaluation.

2. Scene Based Context (SBC) Model

To provide the necessary analysis of SBC models we pick a representative formulation of Gist. To stay consistent with the original work, we will use the same notation as in [12].

Consider an image with image statistics represented by some measurement $\mathbf{v}$. In particular, let $\mathbf{v} = \{\mathbf{v}_L, \mathbf{v}_C\}$, where $\mathbf{v}_L$ refers to statistics in the local spatial neighborhood, at scale σ, around some interest point at location x; $\mathbf{v}_L = \{\sigma, x\}$. $\mathbf{v}_C$ captures the image statistics from the rest of the image (contextual information); $\mathbf{v}_C$ is a low dimensional holistic representation that encodes the structural scene information. In other words, there is a *correlation* between low level representation of the scene and the objects that can be found inside. A typical appearance based object likelihood function $p(O|\mathbf{v}) = \frac{p(O,\mathbf{v})}{p(\mathbf{v})}$, with O being the object of interest, can now be re-written as $p(O|\mathbf{v}) = p(O|\mathbf{v}_L, \mathbf{v}_C)$. It is important to note that majority of the existing approaches to recognition simply omit $\mathbf{v}_C$, and only compute $p(O|\mathbf{v}_L)$. To formally include the contextual information into the objective function, we use Bayes' rule to re-write (1):

$$\begin{aligned} p(O|\mathbf{v}) = \frac{p(O,\mathbf{v})}{p(\mathbf{v})} &= \frac{p(\mathbf{v}_L|O,\mathbf{v}_C)p(O|\mathbf{v}_C)}{p(\mathbf{v}_L|\mathbf{v}_C)} \\ &= \frac{p(\mathbf{v}_L|O,\mathbf{v}_C)p(O|\mathbf{v}_C)}{p(\sigma,x|\mathbf{v}_C)} \\ &= \frac{p(\mathbf{v}_L|O,\mathbf{v}_C)p(O|\mathbf{v}_C)}{p(\sigma|x,\mathbf{v}_C)p(x|\mathbf{v}_C)}, \end{aligned} \quad (1)$$

where $p(\mathbf{v}_L|O,\mathbf{v}_C)$ refers to the spatial relationship between objects: knowing the object label O, and the context of the scene $\mathbf{v}_C$, what is the most probable location of the object in such an image; $p(\sigma|x,\mathbf{v}_C)p(x|\mathbf{v}_C)$ is the normalization term referring to the distribution of scales and locations for various contexts; and finally $p(O|\mathbf{v}_C)$ is the contextual object recognition term.

Let us concentrate on $p(O|\mathbf{v}_C)$. The object label O incorporates the scale at which the object is found, the label, and the location in the image: $O = \{\sigma, o, x\}$. The function of interest here, $p(O|\mathbf{v}_C)$, can thus be factored as:

$$p(O|\mathbf{v}_C) = p(\sigma|x,o,\mathbf{v}_C)p(x,|o,\mathbf{v}_C)p(o|\mathbf{v}_C), \quad (2)$$

where $p(\sigma|x,o,\mathbf{v}_C)$ is the scale selection component, $p(x|o,\mathbf{v}_C)$ is the focus of attention (i.e., the most likely location for the object of interest) and $p(o|\mathbf{v}_C)$ is the contextual priming. This function is further evaluated in [12]. Here, however, by the chain rule of conditional probability, $p(O|\mathbf{v}_C)$ can be decomposed in a number of different ways. For example:

$$p(O|\mathbf{v}_C) = p(o|\sigma,x,\mathbf{v}_C)p(\sigma|x,\mathbf{v}_C)p(x|\mathbf{v}_C), \quad (3)$$

where $p(o|\sigma, x, \mathbf{v}_C)$ is the contextual priming given context, object location and scale, $p(\sigma|x, \mathbf{v}_C)$ is the scale parameter, and $p(x|\mathbf{v}_C)$ determines the most probable location of the object in the image.

In turn, let's examine $p(o|\sigma, x, \mathbf{v}_C)$ in detail. The label of the object is dependent on its physical properties (σ and x), and its surroundings ($\mathbf{v}_C$). Furthermore, it is generally true that physical properties of objects are independent of context: $(x, \sigma) \perp \mathbf{v}_C$. For example, a human face may be of different sizes and may appear in different locations in the image, independent of the context that it is in. Therefore, it is reasonable to assume that if scale and position are independent of context given the object label, then $p(\sigma, x, \mathbf{v}_C|o) = p(\sigma, x|o)p(\mathbf{v}_C|o)$. In turn, $p(o|\sigma, x, \mathbf{v}_C) = \frac{p(o|\sigma,x)p(o|\mathbf{v}_C)}{p(o)}$, since $p(o)$ is constant (i.e., same number of training images per category), this term is omitted for clarity. Thus, we can re-write (2) as follows:

$$\begin{aligned} p(O|\mathbf{v}) &= \frac{p(\mathbf{v}_L|O, \mathbf{v}_C)p(o|\sigma, x, \mathbf{v}_C)p(\sigma|x, \mathbf{v}_C)p(x|\mathbf{v}_C)}{p(\sigma|x, \mathbf{v}_C)p(x|\mathbf{v}_C)} \\ &= p(\mathbf{v}_L|O, \mathbf{v}_C)p(o|\sigma, x)p(o|\mathbf{v}_C). \end{aligned} \quad (4)$$

For the multi object case

$$\begin{aligned} p(o_n|\mathbf{v}_C) &= \sum_{i=1}^{k} p(o_n|C_i, \mathbf{v}_C)p(C_i|\mathbf{v}_C) \\ &\approx \sum_{i=1}^{k} p(o_n|C_i)p(C_i|\mathbf{v}_C), \end{aligned} \quad (5)$$

where k is the number of possible scenes, C_i are various scene context categories, and o_n is the label for the nth object. Finally:

$$p(O_n|\mathbf{v}) = p(\mathbf{v}_L|O_n, \mathbf{v}_C)p(o_n|\sigma, x) \sum_{i=1}^{k} p(o_n|C_i)p(C_i|\mathbf{v}_C). \quad (6)$$

In this approach, the statistics of the local neighborhood $\mathbf{v}_L$ and the contextual information $\mathbf{v}_C$ are both represented using global image features. In particular, in the scene representation proposed in [7], the image is first decomposed by a bank of multiscale oriented filters (tuned to eight orientations and four scales). Then, the output magnitude of each filter is averaged over 16 non-overlapping windows arranged on a 4 grid. The resulting image representation is a $4 \times 8 \times 16 = 512$ dimensional vector. The final feature vector, used to represent the entire image, is obtained by projecting the binned filter outputs onto the first 80 principal components computed on a large dataset of natural images.

Now, as mentioned earlier, another approach to contextual object recognition is possible. In the next section we discuss such an alternative method based only on interactions between individual object labels in the image.

3. Object Based Context (OBC) Model

As a representative of OBC approaches, we selected CoLA, the context-based object recognition system described in [2] based on Co-occurrence, Location and Appearance. To stay consistent with the original work, we will use the same notation as in [2].

At a high level, this representation is built on considering multiple stable segmentations for the input image, resulting in a large collection of segments, though variants also exist using, for example, random segmentations or bounding boxes. Each segment is considered as an individual image and is used as input into a Bag of Features (BoF) model for recognition. Each segment is assigned a list of candidate labels, ordered by confidence. The segments are modeled as nodes of a Conditional Random Field (CRF), where location and object co-occurence constraints are imposed. Finally, based on local appearance and contextual agreement, each segment receives a category label.

3.1. Appearance

As the CoLA approach relies on segmentation based recognition, segment appearance is quantified as in [8]. To review, segments are classified based on a simple nearest neighbor rule with the un-normalized distance of the test segment S_q to class c as:

$$d(S_q, c) = \min_i d(S_q, I_{ic}) = \min_i \|\phi(S_q) - \phi(I_{ic})\|_1. \quad (7)$$

Segment S_q is assigned to its closest category $c_1(S_q)$:

$$c_1(S_q) = \operatorname*{argmin}_c d(S_q, c). \quad (8)$$

Similarly, the S_q is assigned to the rest of the categories: $c_i(S_q) = \text{sort}(d(S_q, c_i)), \forall\ 1 \leq i \leq n$, with sorting in ascending order of distance. In order to construct a probability distribution over category labels for image query segment, we introduce the following definition:

$$p(c_i|S_q) = \left[1 - \frac{d(S_q, c_i)}{\sum_{j=1}^{n} d(S_q, c_j)}\right], \quad (9)$$

and is proportional to the nearest neighbor distance between the query segment S_q and the category: $d(S_q, c)$.

3.2. Location and Co-Occurrences

To incorporate a complete notion of visual context, both spatial and semantic (co-occurrence of labels) contexts must be included into the recognition system. A CRF is used to learn the conditional distribution over the class labeling given an image segmentation. Here, the CRF formulation uses a fully connected graph between a small number of segment labels instead of a sparsely connected

graph on the huge set of all pixels, which yields a much simpler training problem.

Context Model. Given an image I, its corresponding segments $S_1, \ldots, S_k$, and probabilistic per-segment labels $p(c_i|S_q)$ (as in [8]), we wish to find segment labels $c_1, \ldots, c_k \in \mathcal{C}$ such that all agree with the segments' content and are in contextual agreement with one other.

This interaction is modeled as a probability distribution:

$$p(c_1 \ldots c_k|S_1 \ldots S_k) = \frac{B(c_1 \ldots c_k) \prod_{i=1}^{k} p(c_i|S_q)}{Z(\phi_0, \ldots \phi_r, S_1 \ldots S_k)}, \quad (10)$$
$$\text{with } B(c_1 \ldots c_k) = \exp\Big(\sum_{i,j=1}^{k} \sum_{r=0}^{q} \alpha_r \phi_r(c_i, c_j) \Big),$$

where $Z(\cdot)$ is the partition function, q is the number of pairwise spatial relations, and α_r is the weighting for each relation. The marginal terms $p(c|S)$, which are provided by the recognition system, are explicitly separated from the interaction potentials $\phi_r(\cdot)$. To incorporate both semantic and spatial context information into object categorization, namely into the CRF framework, context matrices are constructed.

Location. Spatial context is captured by co-ocurrence matrices for each of the four pairwise relationships (above, below, inside and around). The matrices contain the frequency among objects labels in the four different configurations, as they appear in the training data. An entry (i, j) in matrix $\phi_r(c_i, c_j)$, with $r = 1, \ldots, 4$, counts the number of times an object with label i appears with an object label j for a given relationship r. For a detailed overview of the location descriptor, we refer the reader to [2].

Co-occurrence Counts. The co-occurrences of category labels is computed directly from the above mentioned spatial co-occurrences matrices as described in Section 3.3. An entry (i, j) in the co-occurrence matrix counts the times an object with label i appears in a training image with an object with label j. The diagonal entries correspond to the frequency of the object in the training set: $\phi_0(c_i, c_j) = \phi'(c_i, c_j) + \sum_{k=1}^{|\mathcal{C}|} \phi'(c_i, c_k)$, where $\phi'(\cdot) = \sum_{r=1}^{q} \phi_r(c_i, c_j)$. Therefore the probability of some labeling is given by the model: $p(l_1 \ldots l_{|\mathcal{C}|}) = \frac{1}{Z(\phi)} \exp\Big(\sum_{i,j \in \mathcal{C}} \sum_{r=0}^{q} l_i l_j \cdot \alpha_r \cdot \phi_r(c_i, c_j) \Big)$, with l_i indicating the presence or absence of label i. For a detailed description of this example OBC model, refer to Chapter 3.

4. SBC vs. OBC: a Comparison

In the previous section we formulated both the SBC and the OBC models in a manner suitable for a direct comparison. In the following section we show that both definitions of context extract the same physical and semantic information from images and training set, yet use it quite differently.

4.1. Differences and Similarities

Let us compare

$$p(O_n|\mathbf{v}) = p(\mathbf{v}_L|O_n, \mathbf{v}_C) p(o_n|\sigma, x) \sum_{i=1}^{k} p(o_n|C_i) p(C_i|\mathbf{v}_C) \quad (11)$$

to

$$p(c_1 \ldots c_k|S_1 \ldots S_k) = \frac{B(c_1 \ldots c_k) \prod_{i=1}^{k} p(c_i|S_q)}{Z(\phi_0 \ldots \phi_r, S_1 \ldots S_k)} \quad (12)$$

term by term.

Spatial Context:

$$p(\mathbf{v}_L|O_n, \mathbf{v}_C) \leftrightarrow \frac{\exp\Big(\sum_{i,j=1}^{k} \sum_{r=1}^{q} \alpha_r \phi_r(c_i, c_j) \Big)}{Z(\phi_1 \ldots \phi_r, S_1 \ldots S_k)}, \quad (13)$$

where $p(\mathbf{v}_L|O_n, \mathbf{v}_C)$ refers to estimating the probability of the local patch $\mathbf{v}_L$ containing the object of interest O_n, given the scene information $\mathbf{v}_C$. In other words, assuming the scene context and object identity, where are the probable locations for the object of interest? Similarly, $\frac{\exp(\sum_{i,j=1}^{k} \sum_{r=1}^{q} \alpha_r \phi_r(c_i, c_j))}{Z(\phi, S_1 \ldots S_k)}$, the spatial component of $\frac{B(c_1 \ldots c_k)}{Z(\phi, S_1 \ldots S_k)}$, estimates approximately the same information. Given all the potential objects in the scene, the probability of each spatial arrangement of objects is calculated. However, instead of estimating the absolute location for each candidate object individually, the relative pairwise locations of all objects are chosen simultaneously.

Appearance:

$$p(o_n|\sigma, x) \leftrightarrow p(c_i|S_q), \quad (14)$$

where $p(o_n|\sigma, x)$ is the likelihood of a particular object being present in a given region of the image (region is defined by scale and location). In turn, $p(c_i|S_q)$ is also the likelihood of a particular object, c_i being present at a particular region of the image, yet here the region is defined by segment S_q.

Semantic (co-occurrence) Context:

$$\sum_{i=1}^{k} p(o_n|C_i) p(C_i|\mathbf{v}_C) \leftrightarrow \frac{\exp\Big(\sum_{i,j=1}^{k} \alpha_0 \phi_0(c_i, c_j) \Big)}{Z(\phi_0, S_1 \ldots S_k)}. \quad (15)$$

Here, $\sum_{i=1}^{k} p(o_n|C_i) p(C_i|\mathbf{v}_C)$ captures the semantic context via the scene information C_i. Once the scene category $p(C_i|\mathbf{v}_C)$ is estimated, the most probable object label, o_n, is chosen from the potential labels in the given scene. Alternatively, $\frac{\exp(\sum_{i,j=1}^{k} \alpha_0 \phi_0(c_i, c_j))}{Z(\phi_0, S_1 \ldots S_k)}$, provides a likelihood of all

possible combinations of objects that the existing segments, $S_1 \ldots S_k$, may be labeled with. Only pairwise relationships between object co-occurrences are learned during training.

As shown above, the SBC and OBC models are analogous in terms of the information and statistics they use to apply contextual reasoning to object recognition. However, as we show next, there are a number of differences between the two models that make the OBC model more attractive and empirically more effective.

4.2. Inference

In estimating quantities 11 and 12, it is crucial to understand the processes of inferring the likelihoods, thresholding, and error propagation. In the case of Gist, one first estimates the scene context $p(C_i|\mathbf{v}_C)$, and subsequently the object label, given the chosen scene $p(o_n|C_i)$, as illustrated in Figure 3(a). In particular, choosing the scene context is critical since it constrains the possible object labels in the image. Inferring an incorrect scene from the context reduces the likelihood of identifying the true object labels, see Figure 5 (3 bottom rows in column (b)). Furthermore, only the scenes that have been predefined or learned in training may be considered for an input image, however, objects that exist in the training set may appear in different configurations (scenes) from those in test images, see Figure 5 (bottom row in column (b)). Thus, the accuracy of identifying the labels of objects that exist in an image is critically dependent on identifying the correct scene label for the image. In turn, scene information also requires learning, and is heavily dependent on the training set or manually defined rules.

Alternatively CoLA, an OBC model, Figure 3(b), employs a simple representation and an efficient algorithm for extracting information from visual input without committing to a scene label in a preprocessing stage. Using the traditional Bayesian likelihood estimation of a particular image region being a given object, $p(c_i|S_q)$, a graphical model selects the particular object labels based on the object category co-occurence and spatial relations according to the training data.

Although scene based context is not required for accurate object recognition with an OBC model, we think that scene-level information is indeed an interesting notion. Using the CoLA formulation, this information can be available as a byproduct, rather than as an input, as in Gist. Once the probability of a given set of object labels, $\frac{B(c_1 \ldots c_k)}{Z(\phi_0 \ldots \phi_r, S_1 \ldots S_k)}$, is determined, that set of labels can be mapped to a particular scene.

4.3. Training

Training is a crucial part of any classification task, and object recognition in particular. The two key aspects pertaining to training data are the level of detail in the data labeling and the training set size. Scene based approaches require a large training set since many examples are needed to capture not only the statistics of the object category, but also its scene context [9, 13]. Furthermore, training data must be labeled with the individual object labels, and also with the scene labels. To our knowledge, the majority of object recognition datasets do not contain scene definitions and moreover, it is not clear how to define the scene context. For example, nearly identical scenes may be identified as either *beach* or *coast*, or even as *shore*. Potentially, word hierarchies such as *WordNet* may be used to resolve such ambiguities, but this adds another layer of complexity to the model. Also, as the number of object categories increases, the number of scenes will likely also increase as well and ambiguities between scenes will also be greater.

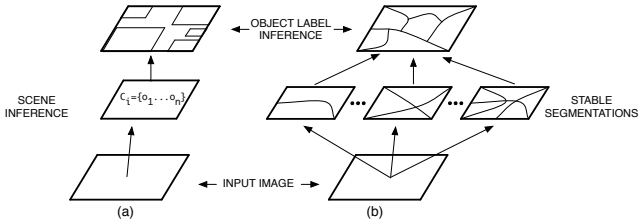

Figure 3. Gist (a) and CoLA (b). Inferring the object labels using Gist requires one first to commit to a scene category and only then infer the object label; with CoLA, no such commitment is necessary.

Approaches based on individual object interactions, however, require considerably less training data as only object appearance and object co-occurrence needs to be learned. In [2, 8] only 30 examples per category were used for training. Only object labels themselves are necessary for training, rather than scene context.

4.4. Scalability

One drawback of the OBC model, is that the required example interactions between object labels are rather sparse in the currently available datasets. Not many object categories co-occur in the same images. However, with the inclusion of many more object categories, the contextual matrices will only get richer and importance of contextual constraints will be even more evident. Note that the complexity of learning co-occurrences is only quadratic in the number of categories since only pairwise relations are computed.

The approach of Gist type methods, which heavily rely on scene information, will perhaps only suffer from an inclusion of additional object categories. New scenes will have to be defined, and the problem of scene inference given the semantic context, $\mathbf{v}_C$ will become even more ambiguous.

5. Empirical Comparison of Contextual Models

In this section we perform an empirical comparison of the two discussed approaches. We used the same subset of the LabelMe, [10], dataset for the experimental comparison as was done by [9]. We trained and tested the CoLA approach with twelve categories. The training set has 15691 images and 105034 annotations and the test set has 560 images and 2026 annotations. The test set comprises images of street scenes and indoor office scenes. To avoid overfitting, street scene images in testing were photographed in a different city from the images in the training set. Figure 5 shows localization and recognition accuracy for example images taken from the LabelMe dataset using Gist and CoLA. Column (c) in Figure 5 shows the accuracy of localization using the stable segmentations used by CoLA. Since this database contains many more categories than just twelve that were chosen by [9], some of the localized regions are not labeled, due to low recognition accuracy, to avoid a forced choice label. In this experiment we mark regions as 'unknown' if the maximum label probability is less than or equal to chance. (On average, of 54 segments per image, 1.51 were labeled as 'unknown'.) Note that the segmentation based approach not only eschews the step of predicting the scene first, thus avoiding as possibly incorrect retrieval set, but it also provides accurate localization with object boundaries rather than bounding boxes. We refer the reader to [9] and [2] for implementation details and runtime complexity for both Gist and CoLA.

The results in Figure 5 show qualitative differences between the two compared models; however, we wish to evaluate the models quantitatively. In Table 1, we report recognition accuracy, true positive rate (TPR), and the false positive rate (FPR) for both models. The results for Gist were taken directly from ROC curves in [9]; results for CoLA are computed from the confusion matrix shown in Figure 4[1]. Since [9] formulated the recognition problem as a detection task, they emphasized the low FPR per bounding box per category, while in recognition problems the TPR is maximized with less attention to FPR. We show TPR rates for the FPR suggested in [9], and show corresponding FPR per image per category, shown in hypercolumn "Gist (low FPR)", rather than per bounding box. TPR and FPR, per image per category, for CoLA are shown in hypercolumn "CoLA". Note that TPR for CoLA is almost 3 fold greater than for Gist, while FPR for CoLA is almost two orders of magnitude lower than that of Gist. This comparison, however, does not isolate the effectiveness of the contextual model itself. In the case of Gist, the underlying detector or classifier (SVM) may be weak, or in the case of CoLA the stable segmentations may be useless. Similar to the work of [8], where the authors show the significant improvement yielded by including of context in the recognition framework, we evaluate the relative improvement of adding context to the Gist method. In Table 1, we show the TPR (at competitive rates) and FPR for the Gist approach with context "Gist (high TPR)", and only the SVM detector module of the "Gist (SVM no context)". Means of both TPR and FPR are within one standard deviation of each other, and the difference between them is not statistically significant. This suggests that recognition rates of the full Gist approach is hindered by its contextual model rather than the underlying detector or classifier. A possible avenue for improvement of the Gist approach could be to entertain multiple scene category hypotheses, rather than committing to the most probable one.

[1]TPR corresponds to the diagonal entries of the confusion matrix and the FPR is the hollow confusion matrix column sum; both refer to the confusion matrix in Figure 4.

	Gist (low FPR)		Gist (high TPR)		SVM (no context)		CoLA	
category	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
tree	9.59%	1.05	76.0%	36.1	53.1%	41.9	78.1%	0.03
building	7.29%	2.09	85.3%	108	60.2%	111	85.8%	0.04
person	21.1%	0.78	68.5%	24.8	78.4%	25.1	64.0%	0.02
sidewalk	7.98%	2.11	70.2%	52.6	66.0%	54.5	74.4%	0.02
car	68.0%	0.03	68.6%	0.83	44.4%	0.89	69.6%	0.03
road	37.0%	0.86	84.6%	31.6	64.3%	29.7	84.7%	0.03
sky	34.5%	1.49	89.6%	106	60.1%	107	91.9%	0.01
motorbike	48.6%	0.81	55.6%	1.19	63.9%	2.10	55.4%	0.02
screen	50.0%	1.17	64.2%	3.81	88.3%	4.57	68.1%	0.02
bookshelf	13.0%	1.04	61.7%	17.9	46.8%	27.8	59.1%	0.03
keyboard	26.5%	0.61	62.0%	10.3	81.4%	15.2	64.5%	0.01
wall	3.08%	0.88	47.7%	84.6	29.2%	61.7	60.0%	0.02
mean	**27.2%**	**1.14**	**63.2%**	**39.9**	**61.4%**	**40.2**	**70.9%**	**0.02**

Table 1. Recognition accuracy (true positive rate TPR) and false positive rate (FPR) per image per category for both Gist and CoLA approaches. **Gist (low FPR):** TPR for the FPR per image per category that was suggested in [9]. **Gist (high TPR):** FPR (from ROC curves in [9]) per image per category for TPR that is comparable to that of CoLA. **SVM (no context):** FPR (also from [9]) per image per category for TPR, without aid of context, that is comparable to one achieved by CoLA. **CoLA:** TPR and FPR per image per category using CoLA. Note that TPR for CoLA is almost 3 fold greater than for Gist (**70.9%** vs. **27.2%**), while FPR for CoLA is almost two orders of magnitude lower than that of Gist (**0.02** vs. **1.14**) per image per category.

6. Discussion

Over the past few years, the role of contextual models has become more prominent in object recognition systems. As the field of contextual object recognition in computer vision evolves, SBC and OBC models have emerged. In the approach proposed by [12], an example of SBC model, contextual information is captured by the statistical summary of the image. This approach may be related to the contextual processing in the human visual system. The SBC model is very intuitive and potentially efficient. An alternative, OBC based, formulation of context for recognition

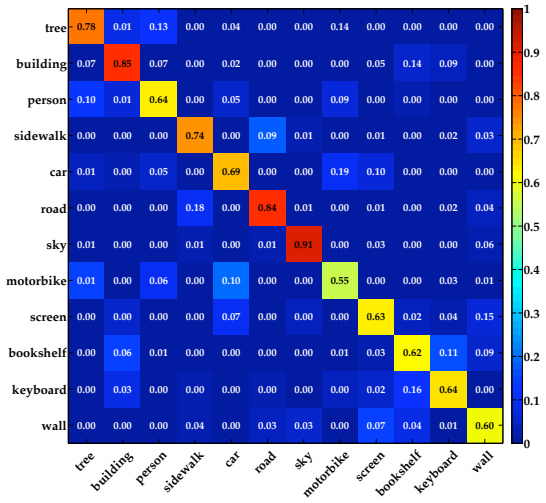

Figure 4. Confusion Matrix for the LabelMe dataset using CoLA.

has recently been proposed. With the OBC model, relationships between individual objects are leveraged, instead of capturing the context of the scene by its low level holistic representation.

Recently, Heitz and Koller of [4], also acknowledged these categories of models and revisited the long standing discussion of *thing vs. stuff* in this context. The authors refer to SBC and OBC as *scene-thing* and *thing-thing* context, respectively. However, neither the SBC nor the OBC models explicitly separate *thing* from *stuff*. In particular, an example OBC model, [2], avoids the *thing/stuff* distinction, and treats all entities to be recognized equally, resulting in *thing-thing/stuff* context. Similarly, instances of SBC models also avoid separation of *thing* and *stuff*, leading to *scene-thing/stuff* context. Perhaps in the future, when *thing vs. stuff* distinction becomes more rigorous, SBC and OBC models will be formulated explicitly using this formalism.

Nonetheless, the comparison of these contextual models shows similarities and differences between them. In particular both models capture analogous physical and semantic information from the image. We demonstrated analytically that the OBC model, although computationally more expensive due to the cost involved in computing the stable segmentations, gives rise to a simpler inference problem. Using the LabelMe database, we empirically compared the two models and showed that CoLA, an approach using an OBC model, considerably outperformed Gist, an SBC based method. The two major differences between OBC and SBC models are the use of stable segmentations vs. sliding window, and the notion of context (object based vs. scene based).

The significant improvement in performance by the OBC model is due in part to the stronger contextual constraints provided by the object-object interactions. But, without a compact representation of image partitions, it is combinatorially difficult to enforce these constraints. Thus, many algorithms tend to settle for scene based contextual connections, which in turn lead to rather confined and weak contextual support. Multiple stable segmentations, in turn, are able to represent the image in such a compact and informative manner for the task of object recognition. Considering thousands of bounding boxes, on the other hand, greatly hinders the false positive rates of the recognition system and leads to intractable inference, as suggested by the experiments. We believe that the shortlist of stable segmentations (aiming for only those segmentations that matter) is the essential substrate for competitive Object Based Context models for object recognition and categorization.

Acknowledgments

This research was supported in part by NSF CAREER Grant #0448615 and NSF IGERT Grant DGE-0333451.

References

[1] M. Fischler and T. Strat. Recognizing objects in a natural environment: a contextual vision system (CVS). *Proceedings of a workshop on Image understanding workshop table of contents*, pages 774–796, 1989. 1

[2] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using Co-Occurrence, Location and Appearance. In *CVPR*, 2008. 2, 3, 4, 5, 6, 7

[3] X. He, R. Zemel, and M. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. *CVPR*, 2004. 1

[4] G. Heitz and D. Koller. Learning Spatial Context: Using Stuff to Find Things. In *ECCV*, 2008. 7

[5] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. *ICCV*, 2005. 1

[6] J. Modestino and J. Zhang. A Markov random field model-based approach to image interpretation. *PAMI*, 1992. 1

[7] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3), 2001. 1, 3

[8] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 2, 3, 4, 5, 6

[9] B. Russell, A. Torralba, C. Liu, R. Fergus, and W. Freeman. Object recognition by scene alignment. In *NIPS*, 2007. 5, 6

[10] B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: a database and web-based tool for image annotation. *MIT AI Lab Memo AIM-2005-025*, 1:1–10, 2005. 6

[11] T. Strat. *Natural object recognition*. 1992. 1

[12] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 2003. 1, 2, 6

[13] A. Torralba, R. Fergus, and W. T. Freeman. Tiny images. Technical report, CSAIL, MIT, 2007. 5

[14] J. Verbeek and B. Triggs. Scene segmentation with crfs learned from partially labeled images. In *NIPS*, 2007. 1

[15] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision (IJCV)*, 2006. 1

(a) (b) (c) (d)

Figure 5. Recognition results for example from LabelMe dataset. (a) Original image. (b) Detected objects by Gist. (c) Recognized objects by CoLA. (d) Ground truth object labeling. *Best viewed in color.*