Exploring Tiny Images: The Roles of Appearance and Contextual Information for Machine and Human Object Recognition

Devi Parikh, *Member, IEEE,* C. Lawrence Zitnick, *Member, IEEE* and Tsuhan Chen, *Fellow, IEEE*

Abstract—Typically, object recognition is performed based solely on the appearance of the object. However, relevant information also exists in the scene surrounding the object. In this paper, we explore the roles that appearance and contextual information play in object recognition. First, through machine experiments and human studies, we show that the importance of contextual information varies with the quality of the appearance information, such as an image's resolution. Our machine experiments explicitly model context between object categories through the use of relative location and relative scale, in addition to co-occurrence. With the use of our context model, our algorithm achieves state-of-the-art performance on the MSRC and Corel datasets. We perform recognition tests, for machines and human subjects, on low and high resolution images, which vary significantly in the amount of appearance information present, using just the object appearance information, the combination of appearance and context, as well as just context without object appearance information (blind recognition). We also explore the impact of the different sources of context (co-occurrence, relative-location and relative-scale). We find that the importance of different types of contextual information varies significantly across datasets.

Index Terms—Object recognition, context, tiny images, blind recognition, image labeling, human studies

1 Introduction

Traditionally, research on recognizing object categories in images has focussed on appearance information pertaining only to the object itself. For instance, parts-based approaches [1], [2] recognize objects by localizing a set of parts corresponding to the local appearance and structure of the object. Popular datasets such as the Caltech datasets [3], [4] have been constructed specifically for such a treatment, where the object to be recognized is found in the center and occupies a significant portion of the image.

In natural images, relevant contextual information about the object also lies in the scene surrounding the object. Recently, many works [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] have attempted to move beyond a purely appearance-based approach by incorporating context using several approaches.

There exist several scenarios, as shown in Fig. 1 in which an object's appearance alone is clearly insufficient for recognition. An example is shown in Fig. 1 (left), where without the context of the rest of the scene (top), it would be hard to recognize the keyboard (bottom). If the amount of intra-class appearance variation is high, or the inter-class appearance variation is low, context may be needed to disambiguate an object's category. For example, as shown in Fig. 1 (center), clothing varies drastically in appearance and is mainly defined by its position relative to the body. In Fig. 1 (right), some object categories such as sky and water, or TV screen and computer monitor have very similar appearance, and may only vary in their relative locations and object surroundings. Other scenarios include those where the amount of appearance information may be limited due to bad image quality, viewing of a scene from a distance, low image resolution, occlusion, etc.

Fig. 1. Illustration of a few scenarios where contextual information is necessary for effective recognition. Left: Impoverished appearance information makes it hard to recognize the keyboard in the image without contextual information; Center: diverse appearance information for the category *clothes* makes it difficult to build a consistent appearance model to describe it; Right: Appearance information is similar for two semantically distinct categories of *TV screen* and *computer monitor* thus requiring contextual information to disambiguate.

In this paper, we explore object level context in the scenario of impoverished image data, where context is *necessary*. Specifically, our goal is object recognition in extremely low resolution images. The need for effective computer vision in low resolution images has many prac-

- *D. Parikh is at Toyota Technological Institute in Chicago (TTIC). E-mail: dparikh@ttic.edu*
- *C. L. Zitnick is at Microsoft Research in Redmond. Email: larryz@microsoft.com*
- *T. Chen is at Cornell University. Email: tsuhan@ece.cornell.edu*

Digital Object Indentifier 10.1109/TPAMI.2011.276

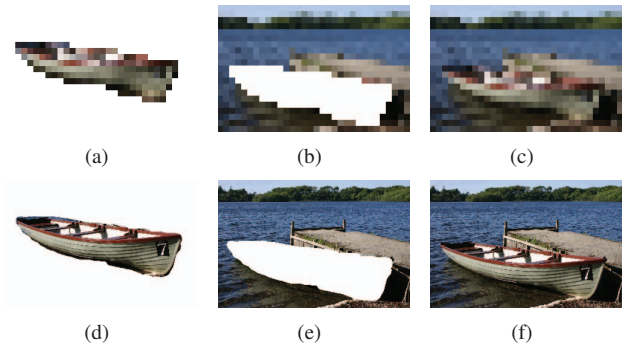

Fig. 2. Example of recognition using appearance alone (a,d), using context alone, i.e. blind recognition (b, e) and context and appearance combined (c, f) for low resolution images (a, b, c) and high resolution images (d, e, f). For low resolution images, context is *necessary* for recognition given the small amount of information provided by the appearance, which is not the case for high resolution. Hence, we advocate exploring context in low resolution images.

tical standings. Low resolution images are space efficient and allow for much faster processing and streaming. Some devices such as older cell phone cameras and web cameras often produce low quality and low resolution images. Images of far away scenes, or images of cluttered complex scenes result in the effective resolution of the individual objects being quite small[1].

Human studies performed in this paper verify that appearance information alone is not enough to accurately recognize objects in low resolution images. However, with the use of context, we find that humans can recognize objects quite reliably, as also observed by Torralba *et al.* [19]. In fact, for the task of blind recognition where appearance information is withheld and only contextual information is given to the subject, recognition accuracy is roughly equal to that of using appearance alone. Through additional human studies we show the relative importance of various types of contextual information, such as co-occurrence, relative location and relative scale information. These studies verify that the task of recognition in low resolution images is an interesting venue for modeling context.

To study the automatic recognition of objects in low resolution images, we propose a segmentation-based approach. Each segment is assigned an object label based on appearance and contextual information learned from a training data set. The beliefs in a segment's labels are computed using a fully connected Conditional Random Field (CRF) with the segments acting as nodes. Context is modeled using the pairwise potentials of the CRF. This formulation allows us to use a wide variety of contextual information, and to compare against human performance in various studies.

Our contributions in this paper are as follows: We perform object recognition in low resolution images; an appropriate scenario for exploring context in which context is *necessary* for accurate recognition. We model context explicitly, and incorporate inter-object relationships in terms of relative location and scale in addition to object co-occurrence. To explore the utility of appearance and contextual information we perform tests on both low and high resolution images, using just object appearance information, using context without object appearance (blind recognition), and the combination of appearance and context, as shown in Fig. 2. These tests were performed both in human and machine experiments. State-of-the-art performances are achieved on the MSRC [20] and Corel [21] datasets. We also explore the roles of each of the different sources of context such as relative location and scale for machine (MSRC and Corel datasets) and human recognition (MSRC and PASCAL [18] datasets), and report some interesting findings.

The rest of the paper is organized as follows: Section 2 outlines some existing related work. Section 3 describes our machine context model. Section 4 describes the experimental set up for our human studies and machine experiments, and provides results and related analysis on the roles of appearance and contextual information in low resolution and high resolution images. Section 5 describes our machine and human-studies experimental set-up and results for exploring the impact of different sources of context for humans and machines. Section 6 raises some interesting points of discussion, followed by a conclusion in Section 7.

2 Related Work

2.1 Context:

Context is believed to play an important role in recognition for humans [22]. Modeling meaningful contextual information for better image understanding has received significant attention in computer vision literature [23], [34]. A variety of information sources may be used to model context. Global scene information, such as global texture [8], [17] or 3D scene information [6] can be used as context. Scene context can be used to restrict the set of possible objects that may be present in the scene, or to reduce the possible locations an object may be present [6], [8], [9], [16], [17].

Context may also be modeled locally. The works of Shotton *et al.* [11] and Fink *et al.* [13] modeled context using local textures, while He *et al.* [10] proposed the use of multi-scale features. Several approaches [10], [11], [12], [14] model short-range interactions for the regularization of region or object labels. The background information surrounding an object has also been proposed for better localization [24]. In our work, we provide human experiments that examine scene contextual information, and local contextual information. For example, subjects may be shown an entire image, or just the pixels contained within a rough bounding box of an object. However, the majority of our work examines the contextual relationships between objects.

1. This is demonstrated in the objects marked 'difficult' in the popular PASCAL visual object category recognition dataset (see Fig. 23).

Numerous works attempt to model the contextual relationships between objects [5], [7], [26], [27]. The early work of Singhal *et al.* [15], used hand-modeled spatial relationships between objects. Torralba *et al.* [7] detect easier to recognize objects first, which in turn aid in the detection of harder objects. Similarly, Heitz *et al.* [31] use easier to recognize textured regions in a scene ('stuff') to better detect objects ('things'). Hoiem *et al.* [6] use 3D information from multiple object types by taking advantage of viewpoint information about the scene. The use of a CRF to enforce co-occurrence relationships between numerous objects was proposed by Rabinovich *et al.* [5] and was later expanded to include spatial relationships [27], [42] and hierarchical models [53]. In our work, we also propose the use of CRFs to model the contextual interactions of objects in our machine experiments. Recently, discriminative models have also been proposed to model the spatial layout of objects [55]. A study of various contextual models for object recognition is provided by Divvala *et al.* [34].

There exists several other areas of research exploring contextual information. An unsupervised approach to learning object relationships is proposed by Parikh *et al.* [28], while Lee *et al.* [32] discovers novel object categories using the context provided by known categories. Gallagher *et al.* [29] and Lin *et al.* [57] propose the use of other types of contextual information, such as social context for analyzing personal photo collections. Yao *et al.* [30] exploit contextual interactions between the human pose and objects for activity analysis. The potential of contextual information can be explored by combining multiple visual sources [33], [34]. While most works leverage context for higher level tasks such as recognition and detection, Parikh *et al.* [25] exploit context for the low-level task of computing saliency maps for images.

2.2 Segmenting objects:

In this paper, we focus on the task of detecting and segmenting objects in a scene using contextual information. Several other approaches have also been proposed for the detection and segmentation task. A pre-computed color-based segmentation of the image may be assigned object labels using appearance and contextual information [5], [27], [42]. Pixel-wise segmentation and detection of objects may be performed by grouping patches using mean-shift [49]. Segmentations can be computed by regularizing the local object labeling of pixels or patches using MRFs [11], aspect models [50] or hierarchical CRFs [53]. Top-down cues for image segmentation are explored by He *et al.* [51]. Gould *et al.* [52] learn the relative locations of objects and use this information to improve upon appearance-based segmentation.

2.3 Low resolution images:

One conclusion of our paper is that the use of context is critical when appearance information is impoverished, such as when images are of low resolution. The use of low resolution images has also been explored by Torralba *et al.* [19] for the recognition of scene categories and object detection using a large database of labeled tiny images. Efros *et al.* [35] recognize human actions in distant videos where the effective resolution of sportsmen is very small.

Human accuracies have been studied in low resolution images for face recognition [36], [37], scene recognition [38], [39], [19], [40] and more recently for object detection [41], [19] and segmentations [19]. However, studies that separate the roles of context from that of appearance as the amount of appearance information varies, and evaluate the impact of the different sources of context, have not been conducted.

3 Approach

In this section, we describe our machine approach to recognizing objects in low resolution images [42]. Descriptions of our human studies, and comparisons to our machine algorithms are given in later sections.

Our goal is to utilize context for recognizing objects in very low resolution images. We obtain these low resolution images by down-sampling images of higher resolution. The aspect ratio of the original image is maintained while reducing the larger dimension to 32 pixels. Torralba *et al.* [19] show that humans can recognize objects in 32×32 images, which our human studies also confirm. Further down-sampling results in a significant degradation in performance [19], [40]. We also apply our method to the original resolution images to study the trade off between appearance and context in different scenarios. The following discussion is common for images of either resolution.

The task we consider is to semantically label every pixel in an image. We approach this task at the region or segment level since good spatial support is shown to significantly help recognition [43], [44]. Hence, our task is to recognize the content of every segment in an image from a pre-determined list of C possible classes. In addition to the appearance information pertaining to the region itself, which we refer to as the data term, we wish to capture the interactions among the different segments through context.

We model this through a fully connected pairwise Conditional Random Field (CRF) similar to [5], where each node corresponds to a segment in the image, and the edges correspond to pair-wise contextual interactions between the segments. In our experiments, the number of segments per image was on average 7 and never exceeded 17, which made such a model feasible. For more complex scenarios containing a larger number of segments, the structure of the graphical model should be intelligently chosen or learnt from data.

We define the conditional probability of our class labels given the segments within our CRF as

$$P(\mathbf{c}|\mathbf{S}) = \frac{1}{Z} \prod_{i=1}^{N} \Psi_i(c_i) \prod_{i,j=1}^{N} \Phi_{ij}(c_i, c_j), \qquad (1)$$

where Z is the partition function. The data term $\Psi_i(c_i)$ computes the probability of class c_i given the appearance of segment $S_i \in \{S_1, \ldots, S_N\}$. The pair-wise potentials

$\Phi_{ij}(c_i, c_j)$ capture the contextual information between segments using co-occurrence statistics from training data at different locations and scales.

3.1 Appearance

Our data term $\Psi_i(c_i) = p(c_i|S_i)$ depends on the texture, shape and color of the segment. To incorporate the texture and shape information, we use the TextonBoost [11] code [45] with one modification. TextonBoost incorporates context through the appearance of surrounding texture patches. Since we are interested in modeling context at the object level and not implicitly through features, we trained TextonBoost on individual objects and not entire images, using the ground truth segmentations. Thus any contextual information captured by TextonBoost from surrounding objects was removed. In our experiments 700 rounds of boosting were performed instead of 5000 as used in [11]. The resulting class likelihoods for each pixel found by TextonBoost are averaged across each segment to obtain a vector with length C equal to the number of possible classes.

To incorporate color, we train a Gaussian Mixture Model (GMM) for each class. We used 7 Gaussians per class in the three-dimensional RGB space. The likelihoods for each pixel are averaged across the segments to obtain a length C vector. In order to combine the results of TextonBoost and the color GMM to obtain $\Psi_i(c_i)$, we use an approach similar to He *et al.* [10]. The two C length vectors are concatenated and passed through a multi-layer perceptron neural network with C outputs. We used 20 hidden layer nodes in our experiments with a sigmoid transfer function.

3.2 Context

The edge-interactions $\Phi_{ij}(c_i, c_j)$ capture the contextual information between segments S_i and S_j through co-occurrence counts given the segments' locations and scales. This is modeled as

$$\Phi_{ij}(c_i, c_j) = [\phi_{ij}(c_i, c_j) + \epsilon]^{\eta}. \quad (2)$$

In all our experiments, ϵ was fixed to be 1 and corresponds to a weak Dirichlet prior. η was 0.02, which controls the effect of context with respect to the data term. Further,

$$\phi_{ij}(c_i, c_j) = \kappa(c_i, c_j)\lambda_{ij}(c_i, c_j)\varphi_{ij}(c_i, c_j), \quad (3)$$

where $\kappa(c_i, c_j)$ captures the likelihood of classes c_i and c_j co-occurring in the image, $\lambda_{ij}(c_i, c_j)$ represents the likelihood of segments S_i and S_j co-occurring at their observed locations given assignments to classes c_i and c_j, and similarly $\varphi_{ij}(c_i, c_j)$ represents the likelihood of segments S_i and S_j co-occurring with their observed scales given assignments to classes c_i and c_j. We describe these next.

3.2.1 Co-occurrence:

$\kappa(c_i, c_j)$ is the empirical probability of classes c_i and c_j co-occurring in an image. This is learnt through MLE counts from the labeled training data.

3.2.2 Location:

We model the location of a segment in an image using a Gaussian Mixture Model with $L = 9$ components. For our experiments the Gaussian means are centered in a 3×3 grid with standard deviations in each dimension equal to half the distance between the means. We define the value $\alpha_l(l_i)$ as the average likelihood of S_i's pixels being in component $l \in L$. Since most images have a horizontal layout we also tried using only 3 bins spaced vertically apart, but the results were significantly worse. The value of $\lambda_{ij}(c_i, c_j)$ is computed as

$$\lambda_{ij}(c_i, c_j) = \sum_{l_i=1}^{L} \sum_{l_j=1}^{L} \alpha_l(l_i)\alpha_l(l_j)\theta_l(l_i, l_j|c_i, c_j), \quad (4)$$

where $\theta_l(l_i, l_j|c_i, c_j)$ are parameters estimated from training data through MLE counts. More specifically, $\theta_l(l_i, l_j|c_i, c_j)$ is the empirical probability of the segments S_i and S_j occurring at locations l_i and l_j given their assignments to classes c_i and c_j. It should be noted that this is a joint distribution, and thus includes both the absolute location and relative location statistics i.e. $\theta_l(l_i, l_j|c_i, c_j)$ combines the information $\theta_l(l_i|c_i)$ and $\theta_l(l_j|l_i, c_i, c_j)$. Since the absolute location is measured relative to the image, the statistic $\theta_l(l_i|c_i)$ can be viewed as contextual information relative to the entire image.

3.2.3 Scale:

The scale is defined as the proportion of the number of pixels in the segment with respect to the number of pixels in the image. As a result, the scale for each segment has a value between 0 and 1. Similar to location, we model the scale using a GMM. The GMM has $K = 4$ components with means evenly spaced between 0 and 1. The standard deviation of the components are set to half the distance between the means. We define $\alpha_s(s_i)$ as the likelihood of a segment belonging to scale s_i. $\varphi_{ij}(c_i, c_j)$ is then computed as

$$\varphi_{ij}(c_i, c_j) = \sum_{s_i=1}^{K} \sum_{s_j=1}^{K} \alpha_s(s_i)\alpha_s(s_j)\theta_s(s_i, s_j|c_i, c_j), \quad (5)$$

where $\theta_s(s_i, s_j|c_i, c_j)$ are parameters estimated from training data through MLE counts. Again, $\theta_s(s_i, s_j|c_i, c_j)$ is the empirical probability of segments S_i and S_j having scales s_i and s_j given their assignments to classes c_i and c_j. As with location, the absolute and relative scale statistics are both captured here.

3.2.4 Inference

We use Loopy Belief Propagation to perform approximate inference on the CRF using a publicly available implementation [46]. After convergence, the label with maximum belief is assigned to the segment. A sampling based inference technique could also be used as in [5].

Using equation (3) we maintain the simplicity of the model proposed in [5], which uses just co-occurrence counts, while capturing richer information through relative

Fig. 3. Low resolution images from the MSRC (top) and Corel (bottom) datasets. The larger dimension is 32 pixels. The objects are often very small, for instance there are only 4 pixels in the faces in the top left image.

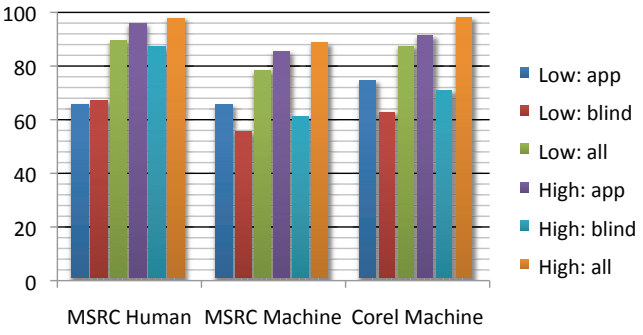

Fig. 4. The recognition accuracies of human subjects and machine on low and high resolution images from the MSRC and Corel datasets using appearance alone (app), blind recognition using context alone (blind) and both appearance and context (all).

location and scale statistics. The proposed model also allows for the straightforward incorporation of additional contextual information, such as relative 3D orientations if available, using the same formulation. We do not do any parameter learning to explicitly increase the likelihood of the training data under our model. Although the current treatment suffices for our purposes, explicit parameter learning such as in [5] may further boost performance.

4 Low Resolution vs. High Resolution

In this paper we present two sets of results on human and machine accuracies. The first set of experiments studies the effect of context on recognition in high resolution vs. low resolution images. In the following section, we present our second set of results studying the use of various types of contextual information.

We study recognition on high and low resolution images using the MSRC dataset [20] and a subset of the Corel dataset [21]. The MSRC dataset contains 591 images with pixel-wise labels coming from 23 classes. Following previous works, we remove 2 classes (horses and mountain) because of very few training instances. The Corel dataset consists of 100 images with labels coming from 7 classes. As stated earlier, we work with images at their original resolution ($\sim 320\times320$) pixels, as well as at low resolution ($\sim 32\times32$ pixels). In both datasets, a random subset of 45% of the images were used for training, 10% for validation and the rest for testing, while maintaining consistent class distributions in these three sets, similar to [11]. We show sample low resolution test images from both datasets in Fig. 3. We first present our machine vision results, followed by a description of our human studies setup and associated results, and finally some analysis of the results obtained.

4.1 Machine Results

For consistency with the human studies (described later), we use the ground-truth segmentations of the images for our first set of experiments (later results use automatic segmentation). We experiment with low and high resolution images, using appearance information alone, contextual information alone (blind recognition) and both appearance and context (entire image). In the appearance-only scenario, the MAP estimates of the data terms were used to label the segments. For blind recognition, the data term corresponding to the segment to be recognized was set to a uniform distribution before running inference on the CRF[2].

The results obtained on the MSRC and Corel datasets are shown in Fig. 4. We use a random subset of 265 images of the MSRC dataset. The results on other random splits are consistent with those shown here. There are several observations we can make. First, the need for context is minimal in the original high resolution images. Appearance alone performs at 86% accuracy on the MSRC dataset, with context increasing performance by 3%. Secondly, appearance provides less information in low resolution images as seen by the drop in accuracy from 86% to 65%. In the scenario of low resolution images, we see that combining appearance and context significantly boosts performance over each individually, to 78% for MSRC and 87% for Corel. It is interesting to note that identical context models were used for images of both resolutions, while the appearance information was trained separately.

We also perform the same experiments with automatic segmentations. We use the Felzenszwalb and Huttenlocher [48] segmentation algorithm (example segmentations are shown in Fig. 5). We find that the use of automatic segmentation does not harm performance significantly. This can be partly attributed to the fact that the training images were also segmented using the same algorithm, resulting in a better match between the training and testing images. Moreover, the ground truth segmentations provided with the MSRC dataset are quite coarse, resulting in the automatic segmentations not being qualitatively very different. Our results are shown in Table 1 along with a comparison to results from previous works when available. In addition to the segment-wise accuracies metric we have used so far, we report pixel-wise accuracies as well. To obtain a pixel-wise label map from our model, all pixels falling within a segment were assigned the segment's predicted label. Only the pixels that were assigned a label in the ground-truth labeling were considered while computing

2. Malisiewicz *et al.* [47] also evaluate their proposed contextual model and other baselines in a blind recognition setting.

TABLE 1
Comparisons of accuracies *

	MSRC		Corel	
	pixel	segment**	pixel	segment
Shotton'06 [11]	58(72)	– (71)	– (75)	–
Yang'07 [49]	62(75)	–	–	–
Verbeek'07 [50]	64(74)	–	–	–
He'04 [10]	–	–	81(80)	–
He'06 [51]	–	–	– (81)	–
Rabinovich'07 [5]	–	– (68)	–	–
Gould'08 [52]	– (77)	–	–	–
Ladicky'09 [53]	75(86)	–	–	–
High	85(91)	84(89)	94(93)	95(93)
Low	81(83)	77(81)	86(86)	85(84)

. * Different splits may have been used for training and testing data
** Segment-wise accuracies may not be directly comparable because the exact settings under which the accuracies were computed may differ

the accuracy (void pixels were ignored). For our own algorithm, we report results on original (high) resolution images that all other works use, as well as on low resolution images. We report average class-wise accuracies, as well as overall accuracies (within parentheses). Even when using low resolution images, our algorithm outperforms previous works on these datasets.

We believe this is due to several reasons. He *et al.* [10] and Shotton *et al.* [11] make decisions at the level of pixels or small patches, while we do so on segments which requires only a few decisions per image. This also allows us to train on segments making the training information more reliable due to inherent aggregation and grouping. Our explicit use of color was found to give a significant boost in performance. A notable observation is that the difference between our average class-wise accuracies and overall accuracy is not very large. Since we model context, we have good performances consistently across categories, including those that have varied appearances and are also less frequent that appearance-based approaches perform poorly on.

4.2 Human Studies Set-up

Our human studies were performed on the MSRC dataset using 11 subjects. The task assigned to them was to identify the outlined segment in the displayed image. We replicate the machine experiments in our human studies. Each subject had to complete two sessions. The first session was on the low resolution images and the second on the original images. In each session, there were three scenarios under which the subjects had to recognize the segments. The first studied appearance-based recognition by only displaying the segment to be recognized without the rest of the image, Fig. 2(a, d). The second studied blind recognition in which the subject was shown the image with the pixels removed from the segment to be recognized, Fig. 2(b, e). The final

Fig. 5. Illustrations of automatic segmentations.

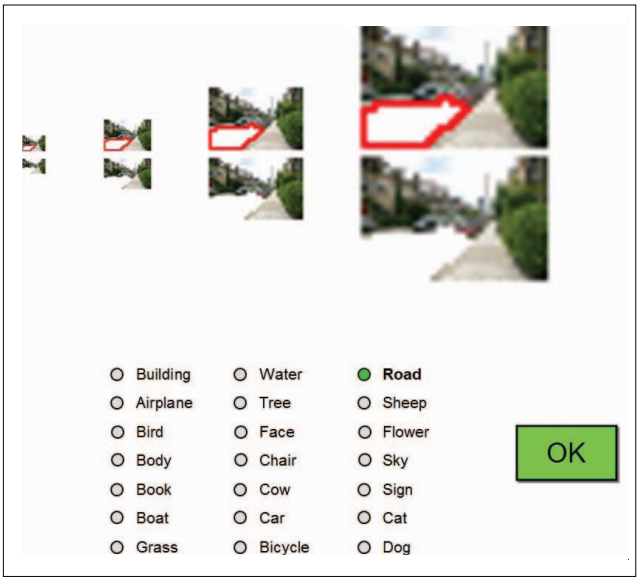

Fig. 6. A snapshot of the interface used for human studies on low resolution images for blind recognition.

scenario displayed the entire image allowing the subject to use both appearance and contextual information for recognition, Fig. 2(c, f). In each scenario the images were displayed with the segment outlined, as well as without the segment outlined to avoid distraction. For low resolution images, the images were displayed at four different scales (32×32, 64×64, 128×128 and 256×256) using bicubic interpolation so that the subjects could focus on whichever scale they desired, without increasing the amount of information being displayed [19]. The list of possible classes from which the subjects could choose was displayed below the images, as shown in Fig. 6. Each subject was asked to recognize 70 segments for each scenario for each resolution (a total of 420 segments per subject). The segments to be recognized were selected randomly from a total of 650 segments in the 265 images per resolution. For consistency, we use the same 265 images of the MSRC dataset for testing as were used in the above machine experiments. On average, subjects took 35 minutes to complete the entire study. The segment boundaries were marked using the ground truth segmentations provided with the MSRC dataset.

4.3 Human Studies Results

The accuracies of the subjects, computed as average class-wise accuracies, are shown in Fig. 4. We see very similar trends in the human numbers as with those from the machine experiments. The need for context is minimal in the original high resolution images. Appearance alone performs at 96% accuracy with context increasing performance by

2%. Secondly, appearance provides less information in low resolution images as seen by the drop in accuracy from 96% to 66%. Interestingly, blind recognition using context alone provides a similar accuracy of 67% for low resolution images. The combination of appearance and context increases accuracy by a statistically significant amount to 89%. This is in agreement with Torralba *et al.*'s observations that human recognition in 32×32 images does not reduce drastically as compared to full resolution images, and we demonstrate here that this is due to inclusion of context. These experiments further support our claim that low resolution images are an interesting venue for modeling context.

It should be noted that the subjects were given a choice of 21 possible category labels. Experiments in which the set of labels is unknown and determined by the subject may yield different results. For some objects the segments are not exact so small amounts of surrounding information, such as grass, may be present for the appearance only tests. Finally, for the task of blind recognition the information inside the segment was removed. However, the rough shape of the segment was still visible and in some cases can supply appearance based information, making the experiment not completely 'blind'. As a result, the accuracies of the blind recognition tests may be artificially high.

4.4 Humans vs. Machine:

Fig. 7 shows the (normalized) confusion matrices of humans and machines on the low resolution images of the MSRC dataset (since the accuracies on the high resolution images are high, we do not show those confusion matrices). While the confusion matrixes show some commonalities, there are significant differences. The four categories from the MSRC dataset that got the highest boost in performance on low resolution images by incorporating context for the human subjects were found to be Body, Face, Water and Boat. The top four categories for the machine were Body, Boat, Building and Sheep, but not Face and Water. This is due to the fact that appearance based recognition for Body and Boat was poor (0% and 30%), having little potential to boost performance of contextually complementary categories such as Water and Face. Moreover, the latter were already reliably recognized (85% and 100%), leaving little room for further improvement.

4.5 High Resolution vs. Context

To compare the category pairs in the human studies that benefited the most from incorporating context to those that benefited the most from incorporating high resolution information, we determine the proportion of the top n (for $n = 50$ & 200) category pairs with most reduction in confusion that are in common between the two. A similar analysis is repeated for the machine experiments. The results obtained can be seen in Fig. 8. We find that the two are in fact correlated, which indicates that the category pairs with low accuracies using low resolution appearance information, can benefit from additional information - be

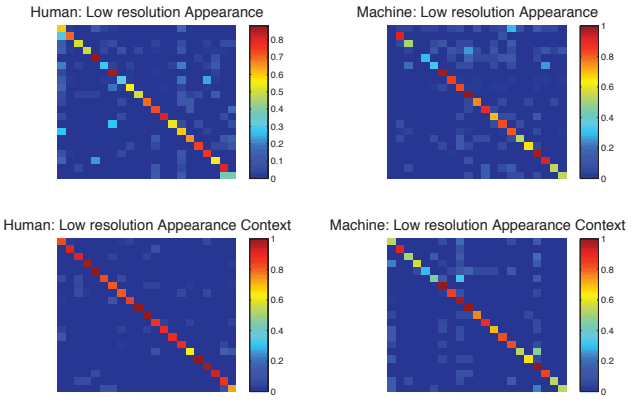

Fig. 7. Confusion matrices of the human studies and machine experiments on the MSRC dataset using the ground truth segmentations (low resolution images).

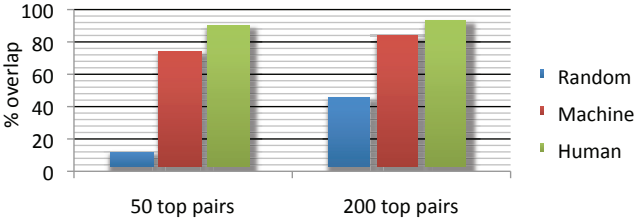

Fig. 8. Evaluating the overlap between pairs of categories that benefitted the most from incorporating context and from incorporating high resolution information, in humans studies and machine experiments.

it in the form of context, or high resolution appearance information. And as our earlier experiments show, once we incorporate high resolution information, context does not provide further boosts in performance. This once again stresses the potential of using low resolution images to model context, as opposed to high resolution images.

4.6 Human Subjects Behavior

We analyzed several aspects of our human studies, that we summarize below:

- The median time taken by subjects to respond to low-resolution images with missing data (appearance alone or context alone) was $\approx$ 4.75 seconds, while that for high resolution images or entire low-resolution images was $\approx$ 2.75 seconds. This is expected since low-resolution images with missing information are ambiguous. It is interesting that the time taken on entire low-resolution images is comparable to that of high resolution images containing contextual information alone.
- If we compare the time taken by subjects, the consistency in responses, and their accuracies (Fig. 4), we see that they all follow a similar trend. Subjects are inaccurate, take longer and are inconsistent among themselves when shown low resolution images with missing information. When shown high resolution images with appearance information alone or entire

images, subjects are accurate, quick and consistent among themselves. We quantify this correlation and indeed find that these quantities are highly correlated (pair-wise correlation coefficient of 0.98).

- We find that the correlation between the average time taken by each of the subjects (averaged across the different scenarios), and their overall accuracy is about -0.4, indicating that how long a subject takes to respond is not an indicator of their accuracy on the tasks. We should note that the difficulty of the tasks themselves has been marginalized out (since we average across the different scenarios).
- We hypothesized that subjects may improve in accuracy as they perform more tests on the same dataset. Surprisingly, we find that the accuracy of the human subjects in the first half of each scenario was not lower than their accuracies in the second half.

5 Different Sources of Context

In this section, we describe human and machine experiments using different sources of context for recognition. These include the use of co-occurrence, relative location and relative scale of objects. In these studies, we use the MSRC [20], Corel [21] and PASCAL 2007 [18] datasets. The MSRC and Corel datasets have shape information available from ground truth segments, where PASCAL 2007 only has object bounding boxes. As a result, the PASCAL dataset does not supply shape information for the task of blind recognition. However, some contextual information may be available in the bounding box when performing appearance-based recognition. The number of categories is similar between the MSRC and PASCAL datasets (21 categories in MSRC, 20 categories in PASCAL 2007), while Corel has 7 categories. The images in PASCAL 2007 are more natural with a large portion of many images not containing any of the 20 objects of interest. Machine experiments were performed on MSRC and Corel, and human experiments were performed on MSRC and PASCAL 2007.

We do not perform machine experiments on PASCAL for a couple of reasons. First, given the poor performance of state-of-the-art techniques at recognizing objects in the PASCAL dataset, attempting the task in 32×32 images is likely to lead to very noisy predictions. Inconclusive results would be generated by the contextual models given the poor initial information. Second, and perhaps more importantly, as also noted recently by Choi *et al.* [54], the PASCAL dataset does not contain interesting contextual interactions. About 50% of the PASCAl images contain only one object (as compared to 22% in MSRC) and 55% of the pixels in the PASCAL images are unlabeled (as compared to 28% in MSRC). Moreover, the PASCAL dataset pre-dominantly consists of people, leading to an entropy of 2.9 of the distribution of object occurrences (as compared to 4.0 in MSRC). It is interesting to view the relative-location contextual statistics for the MSRC and PASCAL datasets (Fig. 9). Bright entries correspond to higher occurrence-statistics. The relative-location statistics

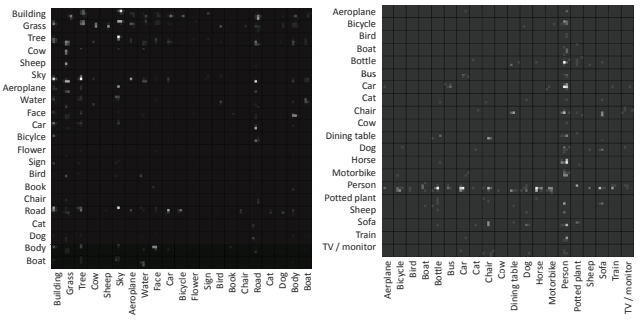

Fig. 9. MSRC (left) and PASCAL (right) relative-location statistics.

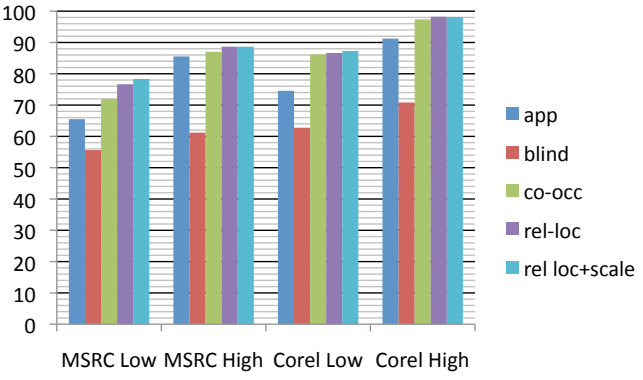

Fig. 10. Machine recognition accuracies on the MSRC and Corel datasets using different sources of context.

are displayed via a 5×5 matrix for each category pair, indicating how often the second category (indicated by the column) is at that particular relative location with respect to the first category (indicated by the row). Overall, we see that the MSRC dataset has more interesting contextual interactions, where as the PASCAL dataset is dominated by the Person category, which co-occur with most other categories in the dataset. We note that these statistics are an artifact of the labeling of the PASCAL images, and not the images themselves (which are more realistic than MSRC). Hence, PASCAL still provides a useful scenario to conduct human studies, which to some extent are not bound by the labels in the images.

5.1 Machine Results

Next, we present analysis on machine experiments using different forms of context (co-occurrence, relative location and relative scale). Average class-wise accuracies using both low and high resolution images from the MSRC and Corel datasets for each of the different forms of context are summarized in Fig. 10. The Corel dataset has fewer classes and the only prominent interactions are the co-occurrence of Polar Bear with Snow, and Rhino/Hippo with Water. Hence, while co-occurrence gives a significant boost in performance on the Corel dataset, relative location and relative scale do not. For MSRC, which is a richer dataset, all forms of context give a significant boost on low resolution images.

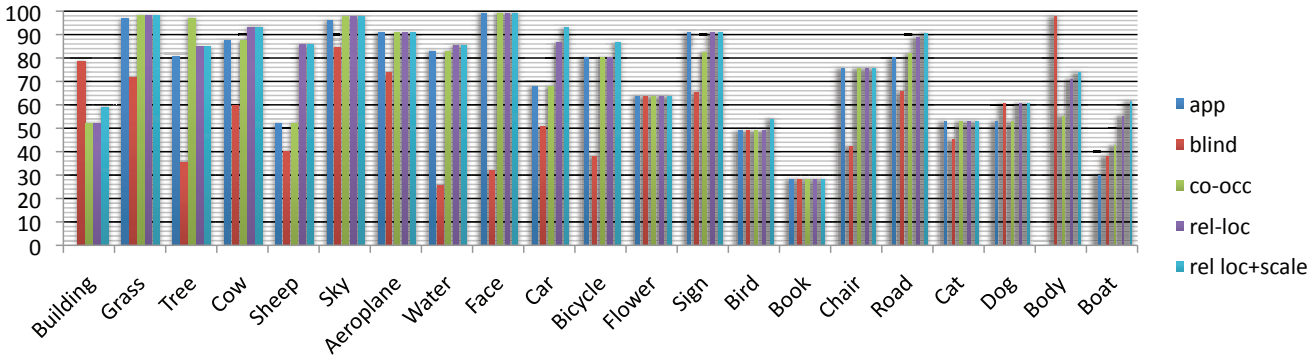

Fig. 11. Average machine accuracies for the 21 categories in the MSRC dataset using appearance alone, using blind recognition with context alone, and using subsequently more complex context models with appearance.

Fig. 12. Images in the MSRC dataset containing books. They occur at similar locations across images, and rarely interact with other categories. Contextual information does not boost the performance of such categories.

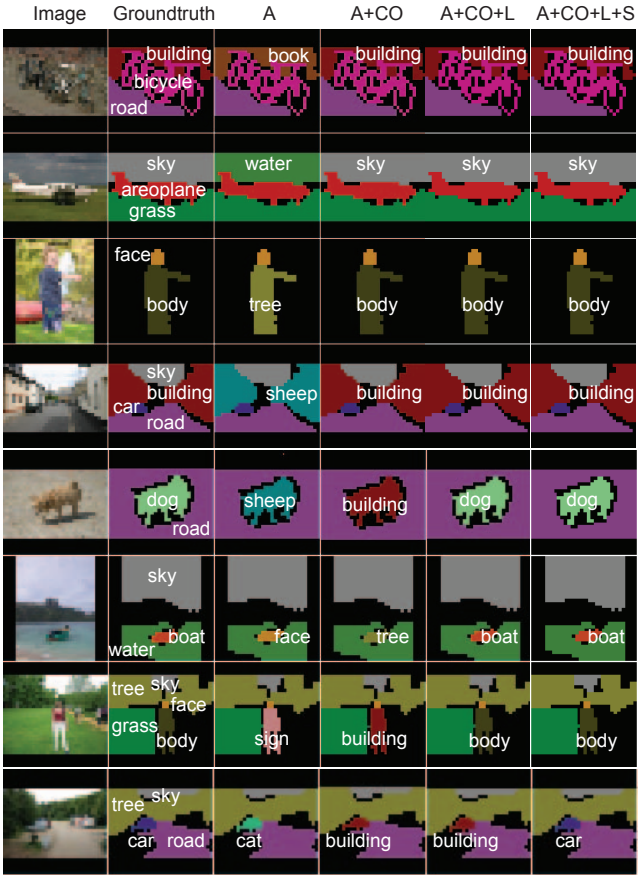

Fig. 13. Illustrations of the effects of different forms of context on recognition. A → appearance, CO → co-occurrence, L → relative location, S → relative scale.

Fig. 11 shows the per class accuracies on low resolution images of the MSRC dataset using only appearance, and subsequently adding the three forms of context. We can see that different object categories benefit from different forms of context. Some categories such as Book and Chairs do not receive any benefit from context due to peculiarities of the dataset, such as they rarely co-occur with other objects (Fig. 12). Categories such as Body and Boat gain significantly from context. Their appearance cues are very weak (0% in the case of Body), but they are very strongly associated with other categories (Face and Water respectively) whose appearance cues are quite reliable. In fact, for some categories such as Body and Building, blind recognition performs much better than appearance information alone as well as combined appearance and context. In several categories, relative scale does not provide a boost in performance. This may be due to lack of scale related dependencies due to inherent semantics of the categories, or due to depth variations of the objects across images, to which our scale measure is not invariant. The independence of scale is automatically learnt by our model. In some categories, albeit rarely, certain forms of context hurt performance. This may be attributed to a category's strong dependence on categories with poor appearance cues. For instance, Sign commonly co-occurs with Building whose appearance term has 0% accuracy.

In Fig. 13 several examples are shown where different types of context helped recognition. Let us consider the last example, where the test image contains Tree, Car, Road and Sky. The appearance alone labels the objects as Tree, Cat, Road and Sky, but the very low likelihood of finding a Cat on the Road along with Tree and Sky made the co-occurrence information flip the label of the Cat to a Building. The location of the Building seems consistent with respect to the Tree, Road and Sky - so the relative location information left the labels untouched. However, the relative scale information discarded the possibility of the Building being so small with respect to the Sky, Tree and Road, and flipped the label of the Building to Car - which matches the ground truth labeling. Examples of incorrect labels provided by the context model are shown in Fig. 14.

Fig. 14. Illustrations of incorrect labelings provided by the context model.

Fig. 15. Low resolution and high resolution appearance information alone (Left: MSRC, Right: PASCAL).

Fig. 16. Co-occurrence information (top: MSRC, bottom: PASCAL).

Fig. 17. Relative-location information (lef: MSRC, right: PASCAL).

5.2 Human Studies Set-up

Next, we study the various sources of context leveraged by humans for recognition. To this end, we perform a series of human studies on Amazon Mechanical Turk using low-resolution images, for which contextual information plays a key role in recognition. In each experiment, the object to be recognized is identified by drawing a red border around it. An image without the red border is also shown in case the red border proves distracting. Unlike the set-up described in section 4.2, we do not display the objects at four different scales. We use a fixed scale that the authors found most comfortable for recognition. We conduct these studies on two datasets: the MSRC dataset containing 625 segments extracted from 265 images and the PASCAL 2007 dataset consisting of 897 objects from 394 images. In order to study the role of different sources of contextual information, we designed the following visualizations using ground truth segmentations (for MSRC) and bounding boxes (for PASCAL).

Appearance: As a baseline, we present only appearance information to the subject, i.e., the object is shown in isolation without additional image information. An example is shown in Fig. 15.

Co-occurrence: We visualize co-occurrence by displaying all labeled objects in the scene side-by-side as shown in Fig. 16. Information about the location and scale of these objects is not available (the objects are scaled to the same size). The object to be recognized is shown with and without a red border, and the remaining objects are shown with and without grey borders.

Relative-location: For relative-location, we display the labeled objects in the same relative locations as they appear in the image, as seen in Fig. 17. All objects are scaled to the same size to remove relative scale information. To avoid overlap of the rescaled objects, the distances between the centers of objects are increased but the relative distances and orientations are kept consistent with the original image.

Relative-scale: The relative-scale visualization is similar to the co-occurrence visualization, but the objects are shown at their true relative scales, i.e. they are not rescaled to the same size. An example can be seen in Fig. 18.

All sources of context combined: We used two visualizations to display relative-location and relative-scale information simultaneously. The first displays all the pixels from labeled objects in the intact image, and the 'void' pixels from unlabeled objects are shown as white. Examples can be seen in Fig. 19 (left), which we call all-no-void. The second visualization (called all-exploded) is shown in Fig. 20 in which additional white space is added between the objects. The relative-location and relative-scale information is available (i.e. all-exploded has the same information as all-no-void), but is similar to the relative-location visualization. This allows us to determine if our choice of visualization affected subjects' accuracies. We note that for PASCAL images, a large portion of the images are often void.

Blind recognition: For sake of completeness and consistency with previous human studies, we also conduct the blind recognition test, where the entire image (including the regions of an image that may be void) is shown, and the pixels belonging to the object of interest are not displayed. Examples can be seen in Fig. 19 (middle). Unlike the MSRC dataset, in the PASCAL dataset the shape information of the object to be recognized is not available and contextual information within the bounding box is lost.

Entire image: We also determine human subjects' accuracy at recognizing objects when the entire image is available. Examples can be seen in Fig. 19 (right). As compared to 'all-exploded' and 'all-no-void', the entire image has two extra sources of information. The first are the void regions of the image that could contain useful contextual information. Second, access to the entire natural image may enable extraction of other sources of information besides relative location and scale, such as 3D geometric contextual cues.

High resolution appearance: Finally, we test human subjects' on recognizing the same objects in high resolution images, without any contextual information, as seen in

Fig. 18. Relative-scale information (top: MSRC, bottom: PASCAL).

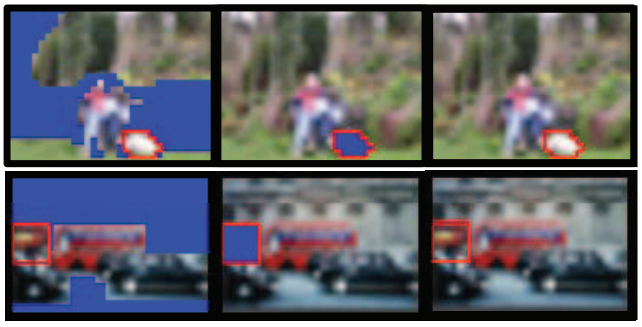

Fig. 19. Left to right: All contextual information via the all-no-void visualization, blind recognition and entire image (top: MSRC, bottom: PASCAL).

Fig. 15.

5.3 Human Studies Results

In this section we present results of the human studies using different sources of context as described above. We obtained responses from 10 subjects on Amazon Mechanical Turk for each test instance. Since the reliability of subjects on Mechanical Turk is variable, we only retain the three responses for each question from the most accurate subjects. We found this calibration step to provide accuracies similar to the authors' accuracies on the same tasks. This may result in accuracies that are artificially high by a small amount. For instance, if the subjects chose their responses randomly and we picked the three highest scores, an accuracy of about 8% would be found. This is slightly higher than the accuracy of random responses without a filtering step ($\approx$ 5%). However, the relative accuracies of our various tests should be consistent.

We now look at the influence of the different sources of context on the human recognition accuracies, as shown in Fig. 21. For the MSRC dataset, we find that co-occurrence and relative-location information provide a boost in performance. However, we see that incorporating the relative-scale information does not provide an improvement in performance over co-occurrence information, or over relative-location information. The choice of visualization, all-exploded or all-no-void, for displaying relative-location and relative-scale information does not affect the subjects' accuracies. We also see that the image regions that are marked as void in the ground truth segmentations do provide useful

Fig. 20. All contextual information via the 'all-exploded' visualization (left: MSRC, right: PASCAL).

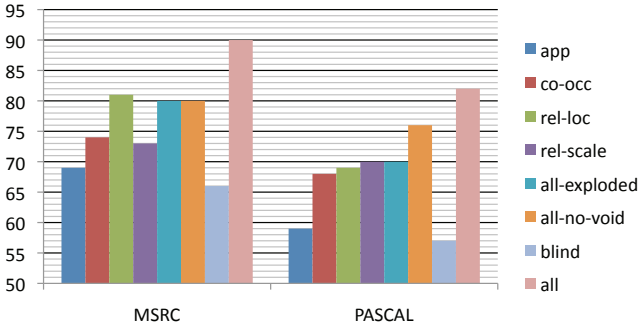

Fig. 21. Recognition accuracies of subjects on the MSRC and PASCAL datasets for different sources of context.

contextual information, which would explain the increase in accuracy from 'all-no-void' to 'all' (entire image)[3].

The PASCAL accuracies are lower overall when compared to the MSRC dataset, especially for appearance information alone and contextual information alone (blind). As with MSRC, we see that relative-scale statistics do not boost recognition performance. Moreover, even relative-location cues seem to be quite weak. This is consistent with our observation that the relative-location statistics of PASCAL have less variation than MSRC, Fig. 9. Interestingly, even though a larger portion of the PASCAL images were marked void, the gap between the accuracies using the entire image, and those using the all-no-void visualization is smaller in PASCAL than in MSRC. This may be because PASCAL has bounding boxes, where the surrounding objects leak into the appearance of the objects, making contextual information less critical overall[4]. To verify this, we conducted an experiment on about 400 low-resolution PASCAL images containing about 800 objects using appearance information alone, comparing performances using segmentations and bounding boxes. We find that using bounding boxes, human accuracy was 51%, while using segmentations, the accuracy was 39%. This indicates that the bounding boxes themselves do in fact incorporate useful contextual information, which is more valuable than the explicit shape information revealed via the segmentations.

The role of the different sources of context for each of the object categories in the MSRC and PASCAL dataset can be seen in Fig. 22. For the MSRC dataset, we see

3. Inspired by this finding, we recently proposed a novel contextual cue that exploits these void regions and boosts performance of a state-of-the-art object detector [58].

4. Previous works have also shown a minimal boost in performance using the PASCAL dataset [55].

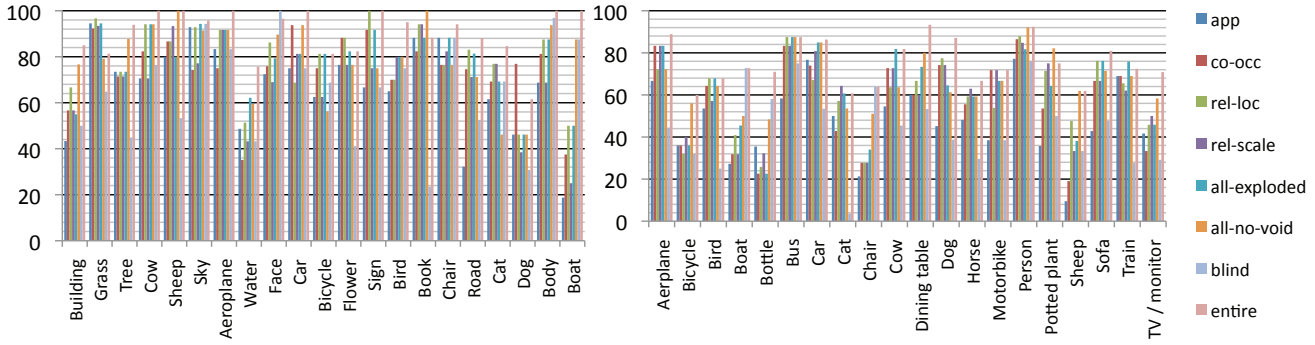

Fig. 22. Recognition accuracies of human subjects on the MSRC (left) and PASCAL (right) datasets for different sources of context on each of the object categories.

some similar trends to the machine results on the roles of context (Fig. 11), where categories such as Body and Boat greatly benefit from contextual information. Unlike machines, humans were able to additionally take advantage of contextual information for categories such as Face and Water, which have complementary categories Body and Boat that are difficult to recognize from appearance information. For sake of completeness we also perform human studies using high resolution appearance information alone, and subjects' accuracies were 97% on both the MSRC and PASCAL datasets. Apart from negligence, some systematic errors made by human subjects in the PASCAL images include scenarios where the object is effectively low res (often seen with Bottle on a dining table which are hard to recognize in isolation), or the bounding box contains two object categories (such as Person on a Bicycle) making it unclear (in spite of instructions) which object should be labeled. In MSRC, apart from low effective resolution of objects, aspects such as nearly white sky, or uncommon view-points of objects (example in Fig. 15 a puppy-dog and sheep can be confused) lead to errors.

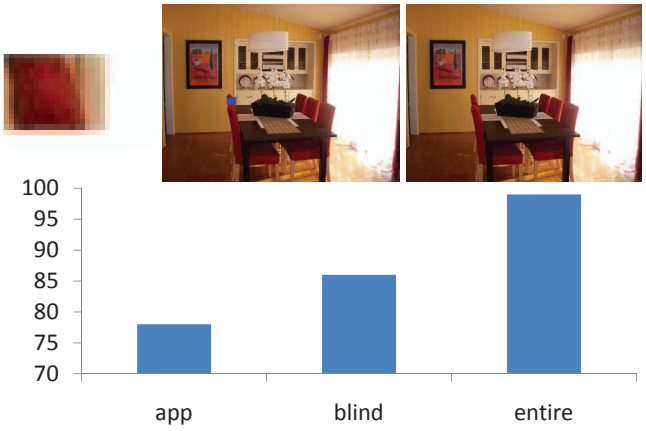

Fig. 23. Human studies on the 'difficult' PASCAL object instances using appearance information alone (left), contextual information alone (blind recognition, middle) and the entire image (right). Similar to object recognition in low-resolution images, appearance information alone is insufficient, and contextual information is necessary for reliable recognition.

6 Discussion

In this section we draw attention to some interesting points of discussion.

Impoverished Appearance Information: As stated in the introduction, low-resolution images are only one scenario where appearance information is impoverished. Other scenarios could include small objects in scenes, occluded objects, etc. Interestingly, the objects marked as 'difficult' in the PASCAL annotations are meant to represent precisely these scenarios. As per the annotation protocol, "an object marked as 'difficult' is considered difficult to recognize, for example an object which is clearly visible but unidentifiable without substantial use of context". These objects are generally ignored in the challenge, but we believe they provide a lucrative venue for exploring contextual information in real-world, natural-occurring images. We perform human studies on 548 PASCAL images containing a total of $1,192$ such 'difficult' object instances, using appearance information alone (Fig. 23 top-left), contextual information alone (blind recognition, Fig. 23 top-middle) and the entire image (Fig. 23 top-right). The accuracies obtained are also shown in Fig. 23 (bottom). Similar to recognition in low resolution images, contextual information is *necessary* in this scenario.

Accuracies on a Dataset by *Chance*

To analyze the amount of contextual information present in a dataset, an interesting metric is to look at what recognition accuracy can be achieved by *chance* as the different forms of context are incorporated. For instance, if we had no information, in a 21 class problem, chance would be $1/21$ i.e. about 5%. However, if we analyze the location statistics of the different categories, and classify a given segment by assigning it to the most likely category given its location, our *chance* accuracy is increased. We still refer to this as chance because no appearance information or other intelligent machinery has been used, we are simply making our best *guess* blindly. Similarly other statistics such as scale, and location and scale combined can be extracted from training data to evaluate what recognition rates can be achieved. This provides some insight into the

significance of the performance boosts achieved by state-of-the-art algorithms.

For the MSRC dataset, we find that we would get 5% recognition rates by chance when classifying each segment/object in the MSRC dataset using uniform prior, 14% using the occurrence-based prior, 30% using the location-based prior and 18% using the scale-based prior. Location and scale based priors combined achieve about 32% accuracy, much higher than the 5% we may be inclined to consider for a 21 class problem. We note that this is the recognition accuracy without looking at a single pixel in the object. Similar analysis of the PASCAL dataset resulted in accuracies of 5% with uniform prior, 48% using occurrence, location and scale information individually and 49% using location and scale both. The high accuracy using occurrence prior alone is, as demonstrated earlier, due to the dominance of the Person category in the statistics of the PASCAL dataset. These results indicate that this holds true across scales and spatial locations, making the PASCAL dataset less interesting for contextual modeling.

Improving Features or Context Models? We explore the question "Do we need to improve our data terms further or our context models to achieve close to human accuracies?" Looking at the MSRC high resolution results in Fig. 4 we find that machines are lagging significantly behind on using appearance information alone. For low resolution images, in which the appearance only tests between humans and machines are similar, the use of context helps humans significantly more. Thus, it appears that improvements in both appearance and contextual models need to be made to match the performance of humans. Since results using only appearance information are similar for humans and machines on low resolution images, this task provides a good scenario for evaluating context models.

Context as Representing the Structure in the World: As we see in our results, the gain from context is certainly a characteristic of the dataset. The more complex a scene, the greater the likelihood of it benefitting from context. As the complexity and number of objects increases, obtaining training datasets with sufficient information will be more difficult. Means of learning context from outside sources such as Google Sets as proposed by Rabinovich *et al.* [5] or large collections of image data such as LabelMe [56] may need to be explored. Leveraging extensive and diverse sources of data is necessary to learn the generic structure of our world, as opposed to potential peculiarities of a dataset.

7 Conclusion

In conclusion this paper makes three main contributions. First, we propose a model for context that includes relative location and scale information, as well as co-occurrence information, which produces state-of-the-art performance on both the MSRC and Corel datasets even with low resolution images. Second, we explore the tradeoffs of appearance and contextual information using both low and high resolution images in human and machine studies. Low resolution images provide an appropriate venue for exploring the role of context, since recognition based on appearance information alone is limited. Finally, we explore the impact of the different sources of context on machine and human object recognition performance in low resolution images, from the MSRC (segment-based) and PASCAL (bounding-box-based) datasets. For human subjects, we find that relative-scale does not prove to be a strong source of contextual information on these datasets, while co-occurrence and relative location are useful.

References

[1] R. Fergus, P. Perona and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *CVPR*, 2003.

[2] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[3] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *CVPR*, Workshop on Generative-Model Based Vision, 2004.

[4] G. Griffin, A. Holub and P. Perona. The Caltech-256 object category dataset. *Caltech Technical Report*, 2007.

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

[6] D. Hoiem, A. Efros and M. Hebert. Putting objects in perspective. *CVPR*, 2006.

[7] A. Torralba, K. Murphy and W. Freeman. Contextual models for object detection using boosted random fields. *NIPS*, 2005.

[8] A. Torralba and P. Sinha. Statistical context priming for object detection. *ICCV*, 2001.

[9] K. Murphy, A. Torralba and W. Freeman. Using the forest to see the trees: a graphical model relating features, objects, and scenes. *NIPS*, 2003.

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

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

[12] P. Carbonetto, N. Freitas and K. Barnard. A statistical model for general contextual object recognition. *ECCV*, 2004.

[13] M. Fink and P. Perona. Mutual boosting for contextual inference. *NIPS*, 2003.

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

[15] A. Singhal, J. Luo and W. Zhu. Probabilistic spatial context models for scene content understanding. *CVPR*, 2003.

[16] B. Bose and E. Grimson. Improving object classification in far-field video. *ECCV*, 2004.

[17] A. Torralba, K. Murphy, W. Freeman and M. Rubin. Context-based vision system for place and object recognition. *AI Memo, MIT*, 2003.

[18] The PASCAL Visual Obejct Classes Challenge. http://www.pascal-network.org/challenges/VOC/voc2007/index.html

[19] A. Torralba, R. Fergus and W. Freeman. 80 Million Tiny Images: A Large Dataset for Non-parametric Object and Scene Recognition. *PAMI*, 2008.

[20] MSRC 21-class Dataset. http://research.microsoft.com/vision/cambridge/recognition/

[21] Corel subset. http://www.cs.toronto.edu/~hexm/label.htm

[22] A. Oliva and A. Torralba. The role of context in object recognition. *Trends Cognitive Science*, 2007.

[23] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 2010.

[24] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 2006.

[25] D. Parikh, C. L. Zitnick and T. Chen. Determining Patch Saliency Using Low-Level Context *ECCV*, 2008.

[26] P. Carbonetto, N. de Freitas and K. Barnard. A statistical model for general contextual object recognition. *ECCV*, 2004.

[27] C. Galleguillos, A. Rabinovich and S. Belongie. Object categorization using co-ocurrence, location and appearance. *CVPR*, 2008.

[28] D. Parikh and T. Chen. Hierarchical semantics of objects (hSOs). *ICCV*, 2007.

[29] A. Gallagher and T. Chen. Estimating age, gender and identity using first name priors. *CVPR*, 2008.
[30] Ba. Yao and L. Fei-Fei. Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. *CVPR*, 2010.
[31] G. Heitz and D. Koller. Learning Spatial Context: Using Stuff to Find Things. *ECCV*, 2008.
[32] Y. J. Lee and K. Grauman. Object-Graphs for Context-Aware Category Discovery. *CVPR*, 2010.
[33] C. Galleguillos, B. McFee, S. Belongie and G. Lanckriet. Multi-Class Object Localization by Combining Local Contextual Interactions. *CVPR*, 2010.
[34] S. Divvala, D. Hoiem, J. Hays, A. Efros and M. Hebert. An Empirical Study of Context in Object Detection. *CVPR*, 2009.
[35] A. Efros, A. Berg, G. Mori and J. Malik. Recognizing action at a distance. *ICCV*, 2003.
[36] T. Bachmann. Identification of spatially queatized tachistoscopic images of faces: how many pixels does it take to carry identity? *European Journal of Cognitive Psychology*, 1991.
[37] L. Harmon and B. Julesz. Masking in visual recognition: effects of two-dimensional noise. *Science*, 1973.
[38] A. Oliva. Gist of the scene. *Neurobiology of Attention, L. Itti, G. Rees and J. Tsotsos (Eds.)*, 2005.
[39] A. Oliva and P. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 1976.
[40] D. Parikh and C. L. Zitnick. The Role of Features, Algorithms and Data in Visual Recognition. *CVPR*, 2010.
[41] D. Parikh and C. L. Zitnick. Finding the Weakest Link in Person Detectors. *CVPR*, 2011.
[42] D. Parikh, C. L. Zitnick and T. Chen. From Appearance to Context-Based Recognition: Dense Labeling in Small Images. *CVPR*, 2008.
[43] T. Malisiewicz and A. Efros. Improving spatial support for objects via multiple segmentations. *BMVC*, 2007.
[44] A. Rabinovich, A. Vedaldi and S. Belongie. Does image segmentation improve object categorization? *Technical Report, UCSD*, 2007.
[45] J. Shotton. http://jamie.shotton.org/work/code.html TextonBoost code.
[46] T. Meltzer. http://www.cs.huji.ac.il/~talyam/inference.html. Inference package for undirected graphical models.
[47] T. Malisiewicz, A. A. Efros. Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships. *NIPS*, 2009.
[48] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.
[49] L. Yang, P. Meer and D. Foran. Multiple class segmentation using a unified framework over mean-shift patches. *CVPR*, 2007.
[50] J. Verbeek and B. Triggs. Region classification with markov field aspect models. *CVPR*, 2007.
[51] X. He. R. Zemel and D. Ray. Learning and incorporating top-down cues in image segmentation. *ECCV*, 2006.
[52] S. Gould, J. Rodgers, D. Cohen, G. Elidan and D. Koller. Multi-Class Segmentation with Relative Location Prior. *IJCV*, 2008.
[53] L. Ladicky, C. Russell, P. Kohli and P. H. S. Torr. Associative Hierarchical CRFs for Object Class Image Segmentation. *ICCV*, 2009.
[54] M. J. Choi, J. Lim, A. Torralba and A. S. Willsky in CVPR 2010 in their paper titled Exploiting Hierarchical Context on a Large Database of Object Categories.
[55] C. Desai, D. Ramanan and C. Fowlkes. Discriminative Models for Multi-Class Object Layout. *ICCV*, 2009.
[56] B. Russell, A. Torralba, K. Murphy and W. Freeman. Labelme: a database and web-based tool for image annotation. *MIT AI Lab Memo*, 2005.
[57] D. Lin, A. Kapoor, G. Hua and S. Baker. Joint People, Event, and Location Recognition in Personal Photo Collections Using Cross-Domain Context. *ECCV*, 2010.
[58] C. Li, D. Parikh and T. Chen. Extracting Adaptive Contextual Cues from Unlabeled Regions. *ICCV*, 2011.

Devi Parikh received her B.S. in Electrical and Computer Engineering from Rowan University in 2005. She received her M.S. and Ph.D. degrees from the Electrical and Computer Engineering department at Carnegie Mellon University in 2007 and 2009 respectively. She was a recipient of the National Science Foundation Graduate Research Fellowship. Since August 2009, she is a Research Assistant Professor at Toyota Technological Institute in Chicago (TTIC), an academic computer science institute affiliated with the University of Chicago. Her research interests include computer vision, pattern recognition and machine learning.

C. Lawrence Zitnick received the PhD degree in robotics from Carnegie Mellon University in 2003. His thesis focused on algorithms for efficiently computing conditional probabilities in large-problem domains. Previously, his work centered on stereo vision, including cooperative and parallel algorithms, as well as developing a commercial portable 3D camera. Currently, he is a researcher at the Interactive Visual Media group at Microsoft Research. His latest work includes object recognition and computational photography. He holds over 15 patents. He is a member of IEEE.

Tsuhan Chen earned his B.S. from National Taiwan University and his M.S. and Ph.D. from Caltech, all in electrical engineering. After working for Bell Labs for several years, he joined the ECE faculty at Carnegie Mellon University (CMU) in 1997. There, in addition to research and teaching responsibilities, he has served as associate department head of ECE and co-director of the Industrial Technology Research Institute (ITRI) Laboratory at Carnegie Mellon, a collaborative research program with ITRI in Taiwan. Since January 2009, he is the director of the School of Electrical and Computer Engineering (ECE) at Cornell University.

In 2007, Chen was elected a fellow of IEEE for contributions in the area of multidisciplinary multimedia signal processing. In 2004 and 2008, he delivered the keynote address at International Conference on Multimedia and Expo (ICME), IEEEs flagship conference on multimedia technologies. From 2002 to 2004, he served as editor-in-chief of the IEEE Transactions on Multimedia, a publication designed to integrate all aspects of multimedia systems and technology, signal processing, and applications.

Chen received the Benjamin Richard Teare Teaching Award in 2006 from the CMU College of Engineering for his consistent excellence in graduate and undergraduate education. Chens success at the intersection of research and education is evident in the awards that he has earned jointly with his graduate students. These include several best paper awards at IEEE Conferences on Computer Vision and Pattern Recognition.

His research group, the Advanced Multimedia Processing (AMP) Lab, has wide interests in various techniques for multimedia applications including computer vision, pattern recognition, computer graphics, multimedia coding and streaming, multimodal biometrics, system implementation and bioinformatics.