From Appearance to Context-Based Recognition: Dense Labeling in Small Images

Devi Parikh
Carnegie Mellon University
dparikh@cmu.edu

C. Lawrence Zitnick
Microsoft Research Redmond
larryz@microsoft.com

Tsuhan Chen
Carnegie Mellon University
tsuhan@cmu.edu

Abstract

Traditionally, object recognition is performed based solely on the appearance of the object. However, relevant information also exists in the scene surrounding the object. As supported by our human studies, this contextual information is necessary *for accurate recognition in low resolution images. This scenario with impoverished appearance information, as opposed to using images of higher resolution, provides an appropriate venue for studying the role of context in recognition. In this paper, we explore the role of context for dense scene labeling in small images. Given a segmentation of an image, our algorithm assigns each segment to an object category based on the segment's appearance and contextual information. We explicitly model context between object categories through the use of relative location and relative scale, in addition to co-occurrence. We perform recognition tests 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 perform these tests in human studies and analyze our findings to reveal interesting patterns. With the use of our context model, our algorithm achieves state-of-the-art performance on MSRC and Corel. datasets.*

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.

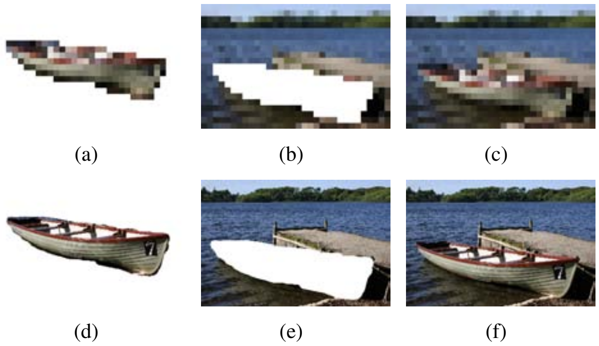

Figure 1: Example of recognition using appearance alone (a,d), using context alone, i.e. blind recognition (b, e) and context and appearance combined (c, f) for low resolution images (a, b, c) and high resolution images (d, e, f). For low resolution images, context is *necessary* for recognition given 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.

In natural images, relevant contextual information about the object also lies in the scene surrounding the object. Recently, many works [5–17] have attempted to move beyond a purely appearance-based approach by incorporating context using several approaches. Global scene information, such as global texture [8, 17] or 3D scene information [6], can be used as context to reduce the set of possible objects that may be present in the scene, or to reduce the possible locations of the objects [6, 8, 9, 16, 17]. Context may also be modeled locally by examining neighboring textures [11, 13], by extracting multi-scale features [10], or by modelling interactions between neighboring regions in the images [10, 12, 14].

Instead of using context to model scene or local texture properties, context may also be used to model higher-level, potentially semantic, interactions among objects [5, 7]. Torralba *et al.* [7] detect easier to recognize objects first, which in turn aid in the detection of harder objects.

Hoiem *et al.* [6] use information from multiple object types by taking advantage of viewpoint information about the scene. Rabinovich *et al.* [5] and Singhal *et al.* [15] proposed the explicit modeling of inter-object context using object co-occurrence, and hand-coded spatial relationships respectively. An unsupervised approach to learning object relationships is proposed by Parikh *et al.* [18].

In many scenarios addressed by prior works context is used to increase recognition accuracy, but it is unclear whether improved use of appearance information could give similar performance boosts. However, there exist several scenarios in which an object's appearance alone is clearly insufficient for recognition. For instance, the amount of appearance information may be limited due to bad image quality, viewing of a scene from a distance, low image resolution, etc. 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, clothing varies drastically in appearance and is mainly defined by its position relative to the body. 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.

In this paper, we explore object level context in the scenario of impoverished image data. Specifically, our goal is dense object labeling in extremely low resolution images. The need for effective computer vision in low resolution images has many practical standings. Low resolution images are space efficient and allow for much faster processing and streaming. Many devices such as 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. 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 images. Efros *et al.* [20] recognize human actions in distant videos where the effective resolution of sportsmen is very small.

As we show in later sections, human studies verify that appearance information alone is not enough for accurate object recognition 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. These studies verify that the task of recognition in low resolution images is an interesting venue for modeling context.

We achieve dense object labeling by assigning labels to a set of pre-computed segments. The segment labels are assigned to be consistent with the contextual information learned from the training data set. The beliefs in a segment's labels are computing 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 for the use of a wide variety of contextual information.

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. These tests were performed both in human and machine experiments. State-of-the-art performances are achieved on the MSRC [21] and Corel [22] datasets.

The rest of the paper is organized as follows. Section 2 describes our context model. Section 3 describes the experimental set up for our human studies and machine experiments, and provides results and related analysis.. Section 4 raises some interesting points of discussion, followed by a conclusion in Section 5.

2. Approach

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]. 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 [23, 24]. 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

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

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.

2.1. 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 [25] 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 C length 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.

2.2. 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.

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.

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.

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

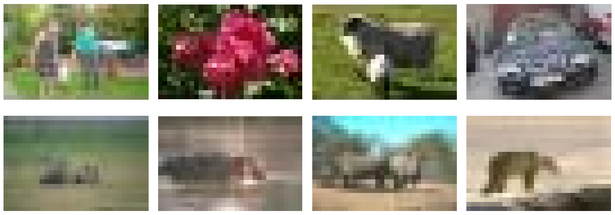

Figure 2: Low resolution images from the MSRC (top) and Corel (bottom) datasets. The larger dimension is 32 pixels. The objects are often very small, for instance there are only 4 pixels in the faces in the top left image.

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.

We use Loopy Belief Propagation to perform inference on the CRF using a publicly available implementation [26]. After convergence, the label with maximum belief is assigned to the segment.

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 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 performances.

3. Results

In our experiments we use the MSRC dataset [21] and a subset of the Corel dataset [22]. 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 ($\sim 320 \times 320$) pixels, as well as at low resolution ($\sim 32 \times 32$ 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 Figure 2. We first describe our human studies, followed by our machine vision experiments and finally some analysis of the results obtained.

3.1. Human Studies

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. 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, Figure 1(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, Figure 1(b, e). The final scenario displayed the entire image allowing the subject to use both appearance and contextual information for recognition, Figure 1(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 Figure 3. 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 265 images (per resolution) from the MSRC dataset. 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.

Human accuracies have been studied in low resolution images for face recognition [27, 28], scene recognition [19, 29, 30] and more recently for object detection and segmentations [19]. However, separating the roles of context from that of appearance as the amount of appearance information varies has not been studied.

The accuracies of the subjects, computed as average class-wise accuracies, are shown in Figure 4 and Table 1. There are several observations we can make. First, the need for context is minimal in the original high resolution images. Appearance alone performs at 96% accuracy with context increasing performance by 2%, which is below statistical significance. Secondly, appearance provides less in-

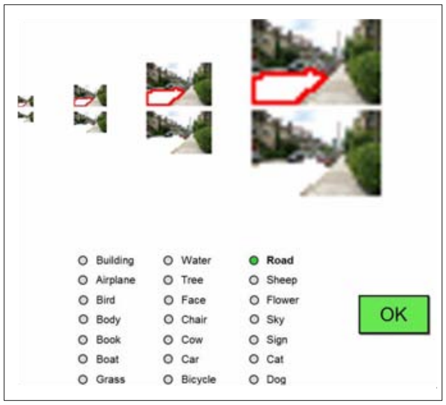

Figure 3: A snapshot of the interface used for human studies on low resolution images for blind recognition.

formation 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 the notion that low resolution images are an interesting venue for modeling context, where the need for context is important.

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. As a result, the accuracies of the blind recognition tests may be artificially high.

3.2. Machine Experiments

We replicate the human studies in our machine experiments. For consistency with the human studies, recognition was performed on the ground truth segmentations (later results use automatic segmentation). 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.

The results obtained on the MSRC dataset are shown in Figure 4 and in Table 1 with results on the Corel dataset. For consistency, we use the same 265 images of the MSRC dataset for testing as were used in the above human studies. The results on other random splits are consistent with those shown here. We see very similar trends in the machine numbers as with those from the human studies. With low resolution images, we see that combining appearance and context significantly boosts performance over each individually, to 78% for MSRC and 87% for Corel. Tests on images with their original resolution show a comparatively smaller, however non-trivial boost in performance. It is interesting to note that identical context models were used for images of both resolutions, while the appearance information was trained separately.

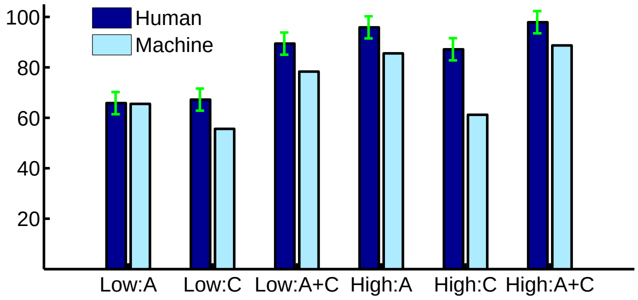

Figure 4: The recognition accuracies of human subjects and machine on low and high resolution images using appearance alone (A), blind recognition using context alone (C) and both appearance and context (A+C). The error bars are also indicated for human accuracies.

Table 1: Machine and human accuracies on MSRC and Corel datasets

	A	C	A+CO	A+CO+L	A + C
MSRC					
Low	65.51	55.62	71.91	76.65	78.33
High	85.55	61.21	87.04	87.73	88.65
MSRC Human					
Low	65.81	67.23	-	-	89.42
High	95.85	87.12	-	-	97.90
Corel					
Low	74.57	62.77	86.19	86.64	87.29
High	91.23	70.84	97.38	98.23	98.16

A → appearance; C → context → co-occurrence CO + relative location L + relative scale

Different sources of context: We present some analysis to evaluate the contribution of the different forms of context (co-occurrence CO, relative location L and relative scale S). Figure 5 shows the per class accuracies on low resolution images using only appearance, and subsequently adding the three forms of context. We can see that different object cat-

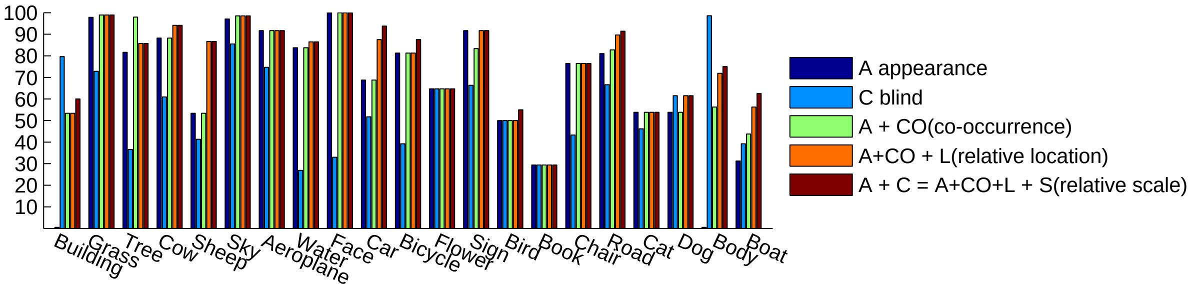

Figure 5: Average accuracies for the 21 categories in the MSRC dataset using appearance alone, using blind recognition with context alone, and using subsequently more complex context models with appearance.

egories benefit from different forms of context. Some categories such as books and chairs do not receive any benefit from context due to peculiarities of the dataset, such as they rarely co-occur with other objects, Figure 6. 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. This lack of dependency 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.

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 Table 1. The Corel dataset has fewer classes and the only prominent interactions are the co-occurrence of polar bears with snow, and rhinos/hippos 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.

In Figure 7 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 re-

Figure 6: 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.

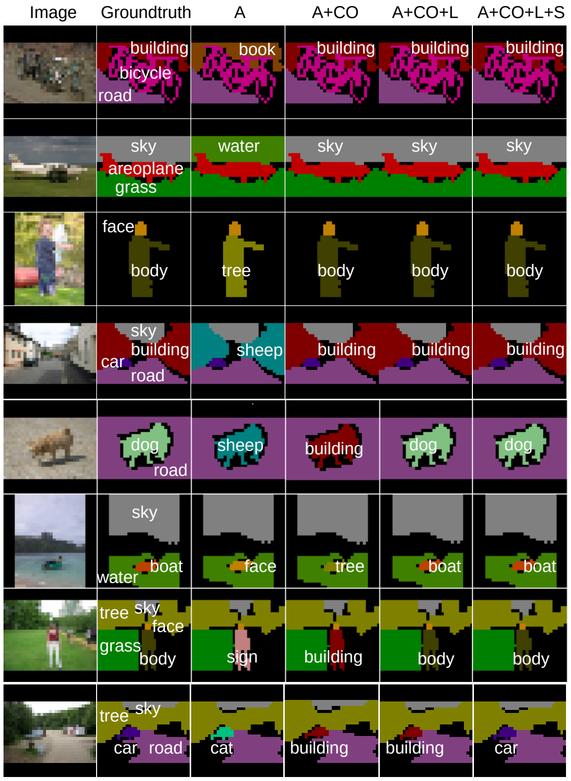

Figure 7: Illustrations of the effects of different forms of context on recognition. A → appearance, CO → co-occurrence, L → relative location, S → relative scale. (Viewed better in color)

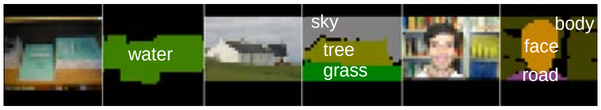

Figure 8: Illustrations of incorrect labelings provided by the context model. (Viewed better in color)

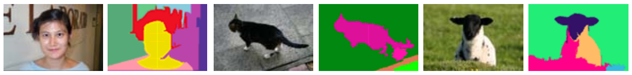

Figure 9: Illustrations of automatic segmentaitons

spect 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. Other intuitive examples are shown in Figure 7 as well. Examples of incorrect labels provided by the context model are shown in Figure 8.

Comparison with other works: We also perform the same experiments with automatic segmentations. We use the Felzenshwalb and Huttenlocher [31] segmentation algorithm (example segmentations in Figure 9). Our results are shown in Table 2 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. 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.

4. Discussion

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

Table 2: Comparisons of accuracies *

	MSRC		Corel	
	pixel	segment**	pixel	segment
[11]	58(72)	– (71)	– (75)	–
[32]	62(75)	–	–	–
[33]	64(74)	–	–	–
[10]	–	–	81(80)	–
[34]	–	–	– (81)	–
[5]	–	– (68)	–	–
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

Humans vs. Machine: We analyze some commonalities and discrepancies between the behavior of humans and machines in incorporating context into recognition. 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 with Body and Face, and Water and Boat being complementary categories. 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 the appearance based recognition for Body and Boat were low (0% and 30%) while Water and Face were very high (85% and 100%), leaving little room for further improvement.

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 Figure 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 improvements on using both appearance and contextual information need to be made to match the performance of humans. Since tests 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 recently proposed by Rabinovich *et al.* [5] or extensive collection of image data such as LabelMe [35] may need to be explored. The easy availability of training data is needed to learn the generic structure of our world, as opposed to potential peculiarities of a dataset.

5. Conclusion

In conclusion this paper contains two main contributions. First, we propose a model for context that includes relative location and scale information, as well as co-occurrence information. Our results show relative location and scale contextual information 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.

In future work, we wish to explore weakly-supervised or even unsupervised learning of the context model, while maintaining its richness. Difficult scenes such as kitchens, offices and streets may require the inclusion of more objects and richer context models.

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

[19] A. Torralba, R. Fergus and W. Freeman. Tiny images. *Technical Report, MIT*, 2007.

[20] A. Efros, A. Berg, G. Mori and J. Malik. Recognizing action at a distance. *ICCV*, 2003.

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

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

[23] T. Malisiewicz and A. Efros. Improving spatial support for objects via multiple segmentations. *BMVC*, 2007.

[24] A. Rabinovich, A. Vedaldi and S. Belongie. Does image segmentation improve object categorization?. *Technical Report, UCSD*, 2007.

[25] J. Shotton. `http://jamie.shotton.org/work/code.html` TextonBoost code.

[26] T. Meltzer. `http://www.cs.huji.ac.il/~talyam/inference.html`. Inference package for undirected graphical models.

[27] 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.

[28] L. Harmon and B. Julesz. Masking in visual recognition: effects of two-dimensional noise. *Science*, 1973.

[29] A. Oliva. Gist of the scene. *Neurobiology of Attention, L. Itti, G. Rees and J. Tsotsos (Eds.)*, 2005. 2

[30] A. Oliva and P. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41:176210, 1976.

[31] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[32] L. Yang, P. Meer and D. Foran. Multiple class segmentation using a unified framework over mean-shift patches. *CVPR*, 2007.

[33] J. Verbeek and B. Triggs. Region classification with markov field aspect models. *CVPR*, 2007.

[34] X. He. R. Zemel and D. Ray. Learning and incorporating top-down cues in image segmentation. *ECCV*, 2006.

[35] B. Russell, A. Torralba, K. Murphy and W. Freeman. Labelme: a database and web-based tool for image annotation. *MIT AI Lab Memo*, 2005.