

Human-Machine CRFs for Identifying Bottlenecks in Scene Understanding

Roozbeh Mottaghi, Sanja Fidler, Alan Yuille, Raquel Urtasun, Devi Parikh

Abstract—Recent trends in image understanding have pushed for scene understanding models that jointly reason about various tasks such as object detection, scene recognition, shape analysis, contextual reasoning, and local appearance based classifiers. In this work, we are interested in understanding the roles of these different tasks in improved scene understanding, in particular semantic segmentation, object detection and scene recognition. Towards this goal, we “plug-in” human subjects for each of the various components in a conditional random field model. Comparisons among various hybrid human-machine CRFs give us indications of how much “head room” there is to improve scene understanding by focusing research efforts on various individual tasks.

Index Terms—Scene Understanding, Semantic Segmentation, Object Detection, Scene Recognition, Human-Machine Hybrid

1 INTRODUCTION

Automatic scene understanding is one of the main goals of computer vision. Given the lofty challenge it presents, the community has historically studied individual tasks in isolation. This includes tasks such as object detection [1], scene recognition [2], contextual reasoning among objects [3], and pose estimation [4]. However, clearly these tasks are related. For example, knowing that the image is a street scene influences where and at what scales we expect to find people. Detecting a microwave in an image can help identify a kitchen scene. Studies have shown that humans can effectively leverage contextual information from the entire scene to recognize objects in low resolution images that can not be recognized in isolation [5].

Recent works [6], [7], [8], [9], have thus pushed on *holistic* scene understanding models. The advent of general learning and inference techniques for graphical models has provided the community with appropriate tools to allow for joint modeling of various scene understanding tasks. These have led to some of the state-of-the-art approaches.

In this paper, we aim to determine the relative importance of the different recognition tasks in aiding scene understanding. We wish to know, which of the tasks if improved, can boost performance significantly. In other words, to what degree can we expect to improve scene understanding performance by improving the performance of individual tasks? We argue that understanding which problems to solve is as important as determining how to solve them. Such an understanding can provide valuable insights

into which research directions to pursue for further improving the state-of-the-art. We use semantic segmentation, object detection and scene recognition accuracies as proxies for scene understanding performance.

We analyze one of the most recent and comprehensive scene understanding models [7]. It is a conditional random field (CRF) that models the interplay between a variety of factors such as local super-pixel appearance, object detection, scene recognition, shape information, class co-occurrence, and compatibility of classes with scene categories. To gain insights into the relative importance of these different factors or tasks, we isolate each task, and substitute a machine with a human for that task, keeping the rest of the model intact. The resultant improvement in performance of the model, if any, gives us an indication of how much “head room” there is to improve performance by focusing research efforts on that task. Note that human outputs are *not* synonymous with ground truth information, because the tasks are performed in isolation. For instance, humans would not produce ground truth labels when asked to classify a super-pixel in isolation into one of several categories. In fact, because of inherent local ambiguities, the most intelligent machine of the future will likely be unable to do so either. Hence, the use of human subjects in our studies is key, as it gives us a *feasible* point (hence, a lower-bound) of what can be done.

Our slew of studies reveal several interesting findings. For instance, we found that human classification of *isolated* super-pixels when fed into the model provides a 5% improvement in segmentation accuracy on the MSRC dataset. Hence, research efforts focused towards the specific task of classifying super-pixels in isolation may prove to be fruitful. Even more intriguing is that the human classification of super-pixels is

• R. Mottaghi is affiliated with AI2. A. Yuille is affiliated with UCLA. S. Fidler and R. Urtasun are with University of Toronto, and D. Parikh is with the Department of Electrical and Computer Engineering at Virginia Tech.

in fact less accurate than machine classification when pixels outside super-pixels are not visible. However when plugged into the model, human potentials provide a significant boost in performance. This indicates that to improve segmentation performance, instead of attempting to build super-pixel classifiers that make fewer mistakes, research efforts should be dedicated towards making the right kinds of mistakes. This provides a refreshing new take on the now well studied semantic segmentation task.

Excited by this insight, we conducted a thorough analysis of the human generated super-pixel potentials to identify precisely how they differ from existing machine potentials. Our analysis inspired a rather simple modification of the machine potentials which resulted in a significant increase of 2.4% in the machine accuracy (i.e. no human involvement) over [7], one of the high performing methods on the MSRC dataset. Also, our human studies and machine experiments on a subset of the PASCAL-Context dataset [10] reveal similar complementary patterns in the mistakes made by humans and machines.

We also studied how well humans can leverage the contextual information modeled in the CRF. We measure human and machine segmentation performance while progressively increasing the amount of contextual information available. We find that even though humans perform significantly worse than machines when classifying isolated super-pixels (with no pixels outside super-pixels visible to the humans), they perform better than machines when both are given access to the contextual information modeled by the CRF.

Our key findings and experiments can be summarized as follows: 1) The performance of humans in super-pixel classification is significantly lower than machines when humans only see the super-pixels and machines use neighboring regions for classification. However, using human super-pixel classification potential inside the CRF model, significantly improves the performance (Section 6.1). 2) Providing contextual information to humans significantly improves their performance, while our machine implementation does not leverage that contextual information as well (Section 7.1). 3) We show that humans have a relatively low performance in recognition when they are shown only super-pixel boundaries (Section 5.4). 4) We propose a new way to estimate contextual co-occurrence of the categories using humans and show that the context statistics obtained by humans are not significantly different from those of the machine (Section 5.2). 5) We show that shape priors obtained from human subjects (drawn by following super-pixel boundaries) are not much better than machine shape priors (Supplementary Material).

Section 2 describes other scene understanding works and also human studies related to computer vision models. In Section 3, we explain the machine

CRF model used for our experiments. In Section 4, the datasets we used for the experiments are described. Section 5 explains how we obtain the machine and human-based potential functions for the CRF model. Section 6 presents the result of plugging-in human or ground truth potential functions in the CRF model. Finally, in Section 7 we analyze the model to see how much potential it holds for improvement and whether the components used in the model are beneficial for humans.

2 RELATED WORK

Scene Understanding: The key motivation behind scene understanding, going back to the seminal work of Barrow in the seventies [11], is that ambiguities in visual information can only be resolved when many visual processes are working collaboratively. A variety of approaches have since been proposed. Many of these works incorporate various tasks in a sequential fashion, by using the output of one task (e.g., object detection) as features for other tasks (e.g., depth estimation, object segmentation) [8], [12], [13], [14], [15]. There are fewer efforts on joint reasoning of the various recognition tasks. In [16], contextual information was incorporated into a CRF leading to improved object detection. A hierarchical generative model spanning parts, objects and scenes is learnt in [17]. Joint estimation of depth, scene type, and object locations is performed in [9]. Spatial contextual interactions between objects have also been modeled [3], [18]. Image segmentation and object detection are jointly modeled in [6], [19], [20] using a CRF. [21] also models global image classification in the CRF. In this paper, orthogonal to these advances, we propose the use of human subjects to understand the relative importance of various recognition tasks in aiding scene understanding.

Human-Studies: Numerous human-studies have been conducted to understand the human ability to segment an image into meaningful regions or objects. Rivest and Cavanagh [22] found that luminance, color, motion and texture cues for contour detections are integrated at a common site in the brain. Fowlkes [23] found that machine performance at detecting boundaries is equivalent to human performance in small gray-scale patches. These and other studies are focused on the problem of unsupervised segmentation, where the task is to identify object boundaries. In contrast, we are interested in scene understanding, including the task of identifying the semantic category of each pixel in the image.

Several works have studied high-level recognition tasks in humans. Fei-Fei *et al.* [24] show that humans can recognize scenes rapidly even while being distracted. Bachmann *et al.* [25] show that humans can reliably recognize faces in 16×16 images, and Oliva *et al.* [26] present similar results for scene recognition. Torralba *et al.* [5] show that humans can reliably detect

objects in 32×32 images. In contrast, we study human performance at tasks that closely mimic existing holistic computational models for scene understanding in order to identify bottlenecks, and better guide future research efforts.

Human debugging i.e. using human subjects to identify bottlenecks in existing computer vision systems has been recently explored for a number of different applications such as analyzing the relative importance of features, amount of training data and choice of classifiers in image classification [27], of part detection, spatial modeling and non-maximal suppression in person detection [28], of local and global image representations in image classification [29], and of low-, mid- and high-level cues in detecting object contours [30]. In this work, we are interested in systematically analyzing the roles played by several high- and mid-level tasks such as grouping, shape analysis, scene recognition, object detection and contextual interactions in *scene understanding*. While similar at the level of exploiting human involvement, the problem, the model, the methodologies of the human studies and machine experiments, as well as the findings and insights are all novel.

This work is an extension of [31]. [31] addresses only the semantic segmentation task. Here we look at two additional tasks: object detection and scene recognition. Additionally, we analyze whether the pipeline is capable of achieving near-perfect accuracies for semantic segmentation, object detection and scene recognition. We also report results on a more challenging dataset.

3 CRF MODEL

We analyze the CRF model of [7] which performs scene understanding by jointly reasoning about a variety of components. While the model shares similarities with past work [19], [21], [32], we choose this model because it provides the best performance among the mentioned methods, and thus forms a great starting point to ask “which components need to be improved to push the performance further?”. Moreover, it has a simple “plug-and-play” architecture making it feasible to insert humans in the model. Inference is performed via message passing [33] and so it places no restrictions (e.g. submodularity) on the potentials. This allows us to conveniently replace the machine potentials with human responses: after all, we cannot quite require humans to be submodular!

We now briefly review this model (Figure 1). We refer the reader to [7] for further technical details. The problem of scene understanding is formulated as that of inference in a CRF. The random field contains variables representing the class labels of image segments at two levels in a segmentation hierarchy: super-pixels and larger segments. To be consistent with [7], we will refer to them as segments and super-segments. The model also has binary variables indicating the

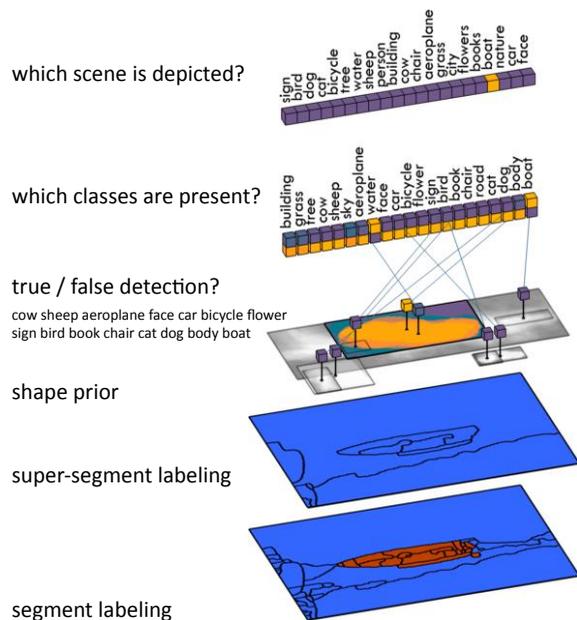


Fig. 1: Overview of the scene model of [7] that we analyze using human subjects. For clarity, not all connections in the model are shown here.

correctness of candidate object detection bounding boxes. In addition, a multi-labeled variable represents the scene type and binary variables encode the presence/absence of a class in the scene.

The segments and super-segments reason about the semantic class labels to be assigned to each pixel in the image. The model employs these two segmentation layers for computational efficiency: the super-segments are fewer but more densely connected to other parts of the model. The binary variables corresponding to each candidate bounding box generated by an object detector allow the model to accept or reject these detections. A shape prior is associated with these nodes encouraging segments to take on corresponding class labels. The binary class variables reason about which semantic classes are present in the image. This allows for a natural way to model class co-occurrences as well as scene-class affinities. These binary class variables are connected to i) the super-segments via a consistency potential that ensures that the binary variables are turned on if a super-segment takes the corresponding class label ii) binary detector variables via a similar consistency potential iii) the scene variable via a potential that encourages certain classes to be present in certain scene types iv) to each other via a potential that encourages certain classes to co-occur more than others.

More formally, let $x_i \in \{1, \dots, C\}$ and $y_j \in \{1, \dots, C\}$ be two random variables representing the class label of the i -th segment and j -th super-segment. We represent candidate detections as binary random variables, $b_i \in \{0, 1\}$, taking value 0 when the detection is a false detection. A deformable part-based model [1] is used to generate candidates. The detector provides us with an object class (c_i), the score (r_i), the

location and aspect ratio of the bounding box, as well as the root mixture component ID that has generated the detection (m_i). The latter gives us information about the expected shape of the object. Let $z_k \in \{0, 1\}$ be a random variable which takes value 1 if class k is present in the image, and let $s \in \{1, \dots, C_l\}$ be a random variable representing the scene type among C_l possible candidates. The parameters corresponding to different potential terms in the model are learnt in a discriminative fashion [34]. Before we provide details about how the various machine potentials are computed, we first discuss the dataset we work with to ground further descriptions.

4 DATASET

We use the standard MSRC-21 [35] semantic labeling benchmark, also used by [7], as it contains “stuff” (e.g., *sky, water*) as well as “things” (i.e., shape-defined classes such as *cow, car*). The PASCAL dataset is more challenging in terms of object (“things”) detection and segmentation, but a large portion of its images, especially “stuff”, is unlabeled. Hence, we augment a subset of PASCAL dataset with additional categories [10] and report results on that.

We use the more precise ground truth of MSRC provided by Malisiewicz and Efron [36] and used in [7], as it offers a more accurate measure of performance. We use the same scene category and object detection annotations as in [7]. Table 1 lists this information. As the performance metric we use average per-class recall (average accuracy). Similar trends in our results hold for average per-pixel recall (global accuracy [19]) as well. We use the standard train/test split from [37] to train all machine potentials, described next.

5 MACHINE & HUMAN CRF POTENTIALS

We now describe the machine and human potentials we employed. Section 6 presents the results of feeding the human “potentials” into the machine model. Our choices for the machine potentials closely follow those made in [7]. For human potentials, we performed all human studies on Amazon Mechanical Turk. Unless specified otherwise, each task was performed by 10 different subjects. Depending on the task, we paid participants 3–5 cents for answering 20 questions. The response time was fast, taking 1 to 2 days to perform each experiment. We randomly checked the responses of the workers and excluded those that did not follow the instructions¹. More than 500 subjects participated in our studies that involved $\sim 300,000$ crowd-sourced tasks, making the results obtained likely to be fairly stable across a different sampling of subjects.

1. As our experiments will demonstrate, the fact that we can train the CRF parameters on responses of human subjects and have it generalize well to human responses to held out test images vouches for the reliability of the collected human responses.

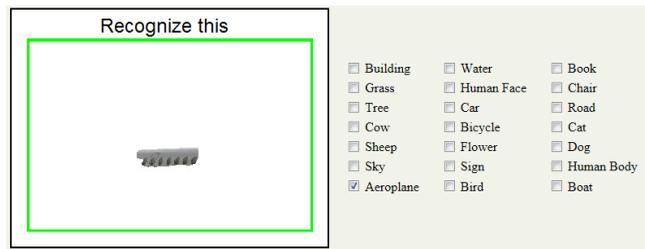


Fig. 2: Segment labeling interface. We ask the human subjects to choose the category that the segment belongs to. If the subjects are confused among a few categories, they have the option of choosing more than one answer.

5.1 Segments and super-segments

Machine: We use UCM [38] to create our segments and super-segments as it returns a small number of segments for each instance of a category. We use thresholds 0.08 and 0.16 for the segments and super-segments respectively. On average, this results in 65 segments and 19 super-segments per image for the MSRC dataset. We use the output of the modified TextonBoost [35] in [32] to get pixel-wise potentials and average those within the segments and super-segments to get the unary potentials. Following [39], we connect these two levels via a pairwise P^n potential that encourages segments and super-segments to take the same label.

Human: Now, we describe how we compute the segment classification potentials for the human case. Of course, ground truth segmentation annotations are themselves generated by humans, but by viewing the whole image and leveraging information from the entire scene. In this study, we are interested in evaluating how each recognition task in *isolation* can help the overall performance.

To follow the exact procedure of TextonBoost, we need to show human subjects a window around each pixel in a segment and have human subjects classify the pixels into one of the semantic categories. Then, we can average the results over pixels to obtain the segment classification potential. However, this procedure is prohibitively expensive as there are ~ 1 million pixels in each image. Therefore, instead of the above procedure, we show all pixels of a segment to subjects and ask them to classify the segment. Hence, just like a machine has a potential for every segment, we have a “human” potential for every segment. This allows us to replace the machine potential with a human potential in the model.

We experimented with several interfaces e.g., showing subjects a collection of segments and asking them to click on all the ones likely to belong to a certain category, or allowing a subject to select only one category per segment, etc. before converging to the one that resulted in most consistent responses from subjects (Figure 2) where subjects are asked to select all categories that a segment may belong to.

Figure 3 shows examples of segmentations obtained

semantic labels													scene types										obj. detection classes																																		
building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	sign	bird	dog	cat	bicycle	tree	water	sheep	person	building	cow	chair	aeroplane	grass	city	flowers	books	boat	nature	car	face	cow	sheep	aeroplane	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat

TABLE 1: MSRC-21 dataset information



Fig. 3: Human labeling results on isolated segments.

by assigning each segment to the class with most human votes. The black regions correspond to either the “void” class (unlabeled regions in the MSRC dataset) or to small segments not being shown to the subjects. Assigning each segment to the class with the highest number of human votes achieves an accuracy of 72.2%. Although usually a 200×200 window around each pixel is used for TextonBoost, we tried different window sizes (10×10 , 20×20 , 30×30 , and 40×40), which resulted in 66.3%, 70.8%, 75.6%, and 77.2% accuracy, respectively². It is interesting that humans perform worse than some machine cases. However, they are better at classifying certain “easy”, distinctive classes or classes they are familiar with e.g., *faces* (see confusion matrix in Figure 11(b)).

The C dimensional human unary potential for a (super)segment is proportional to the number of times subjects selected each class, normalized to sum to 1. We set the potentials for the unlabeled (smaller than 500 pixels) (super)segments to be uniform.

5.2 Class occurrence and co-occurrence

Machine: We use class-occurrence statistics extracted from training data as a unary potential on z_k . We also employ pairwise potentials between z_i and z_k that capture co-occurrence statistics of pairs of classes. However, for efficiency reasons, instead of using a fully connected graph, we use a tree-structure obtained via the Chow-Liu algorithm [40] on the class-class co-occurrence matrix.

Human: To obtain class-occurrence, we showed subjects 50 random images from the MSRC dataset to help them build an intuition for the image collection (not to count the occurrence of objects in the images). For all pairs of categories, we then ask subjects which

category is more likely to occur in an image from the collection. We build the class unary potentials by counting how often each class was preferred over all other classes. We ask MAP-like questions (“which is more likely”) to build an estimate of the marginals (“how likely is this?”) because asking subjects to provide scalar values for the likelihood of something is prone to high variance and inconsistencies across subjects.

To obtain the human co-occurrence potentials we ask subjects the following question for all triplets of categories $\{z_i, z_j, z_k\}$: “Which scenario is more likely to occur in an image? Observing (z_i and z_j) or (z_i and z_k)?”. Note that in this experiment we did not show subjects any images. The obtained statistics thus reflect human perception of class co-occurrences as seen in the visual world in general rather than the MSRC dataset. Given responses to these questions, for every category z_i , we count how often they preferred each category z_j over the other categories. This gives us an estimate of $P(z_j|z_i)$ from humans. We compute $P(z_i)$ from the training images to obtain $P(z_i, z_j)$, which gives us a 21×21 co-occurrence matrix. We use the Chow-Liu algorithm on this matrix, as was used in [7] on the machine class co-occurrence potentials to obtain the tree structure, where the edges connect highly co-occurring nodes. As shown in Figure 4, the structure of the human tree is quite similar to the tree obtained from the MSRC training set. For example, in both trees, there are edges between *grass* and categories like *cow*, *sheep*, and *flower*. However, some edges exist in the human tree that are missing in the machine tree e.g., the edge between *sky* and *bird*.

5.3 Detection

Machine: Detection is incorporated in the model by generating a large set of candidate bounding boxes

². This accuracy is calculated only over segments larger than 500 pixels that were shown to humans.

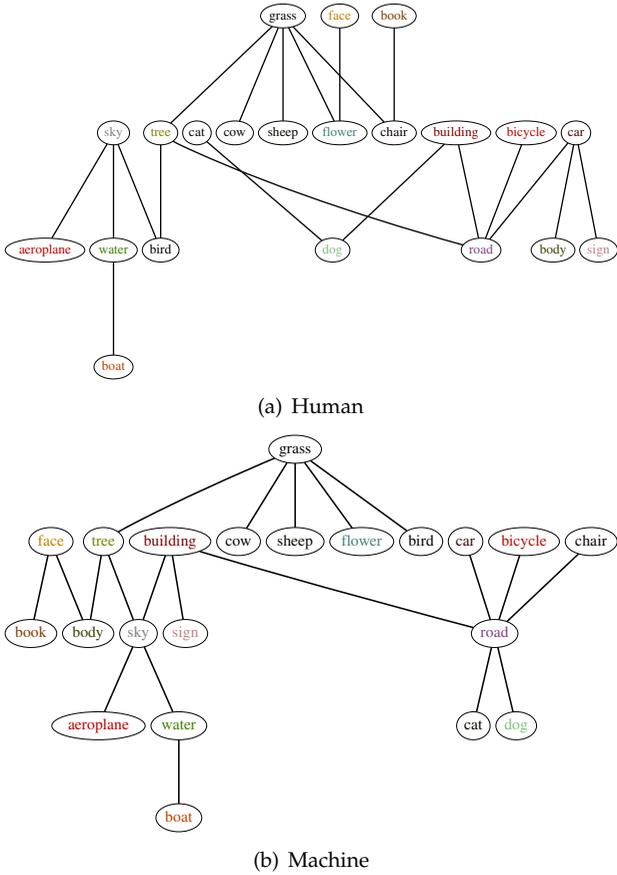


Fig. 4: Chow-Liu trees for humans and machine. The trees share several similarities.

using the deformable part-based model [1] which has multiple mixture components for each object class. The CRF model reasons about whether a detection is a false or true positive. On average, there are 16 hypotheses per image. A binary variable b_i is used for each detection and it is connected to the binary class variable, z_{c_i} , where c_i is the class of the detector that fired for the i -th hypothesis.

Human: Since most objects in the MSRC dataset are quite big, it is expected that human object detection would be nearly perfect. As a crude proxy, we showed subjects images inside ground truth object bounding boxes and asked them to recognize the object. Performance was almost perfect at 98.8%. Hence, we use the ground truth object bounding boxes to simulate human responses.

5.4 Shape

Machine: Shape potentials are incorporated in the model by connecting the binary detection variables b_i to all segments x_j inside the detection’s bounding box. The prior is defined as an average training mask for each detector’s mixture component. The values inside the mask represent the confidence that the corresponding pixel has the same label as the detector’s class. In particular, for the i -th candidate detection,

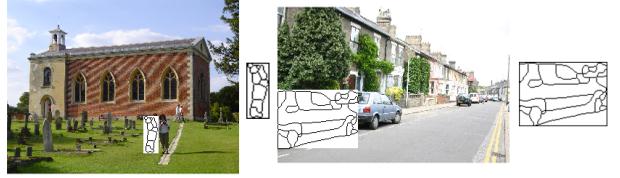


Fig. 5: Human object recognition from image boundaries. We show subjects segments inside the object bounding box and ask them to recognize the category of the object. We show the segments with (left image) and without (right image) context.

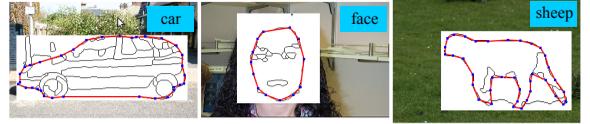


Fig. 6: Human shape mask labeling interface. Human subjects were asked to draw the object boundaries along the segment contours.

this information is incorporated in the model by encouraging the x_j segment to take class c_i with strength proportional to the average mask values within the segment. In the supplementary material, we explain different shape priors that we have used.

Human: We showed 5 subjects the segment boundaries in the ground truth object bounding boxes along with its category label and contextual information from the rest of the scene. Figure 6 shows a few examples. We showed subject contextual information around the bounding box because without it humans were unable to recognize the object category reliably using only the boundaries of the segments in the box (55% accuracy). With context, classification accuracy was 94%. See Figure 5 for example images.

Using the interface of [41], subjects were asked to trace a subset of the segment boundaries to match their expected shape of the object. The accuracy of the best of the 5 masks obtained for each object (normalized for foreground and background) was found to be 80.2%. The best automatic accuracy we obtain with the machine is 78.8% using the distance transform approach (refer to the supplementary material), not much worse than the human subjects’ accuracy. This shows that humans can not decipher the shape of an object from the UCM segment boundaries much better than an automatic approach. Clearly, the UCM segment boundaries are not any more informative to humans than they are to machines.

5.5 Scene and scene-class co-occurrence

Machine: We train a classifier [2] to predict each of the scene types, and use its confidence to form the unitary potential for the scene variable. The scene node connects to each binary class variable z_i via a pairwise potential which is defined based on the co-occurrence statistics of the training data, i.e., likelihood of each class being present for each scene type.



Fig. 7: Human scene classification. Subjects were shown images at multiple resolutions. Subjects were asked to choose the scene category for each image.



Fig. 8: Average scenes for some example scene categories in MSRC.

Human: To obtain scene unary, we ask human subjects to classify an image into one of the 21 scene categories used in [7] (see Table 1). Images were presented at varying resolutions (i.e., original resolution, smallest dimension rescaled to 32, 24 and 20 pixels) as shown in Figure 7. Subjects were allowed to select more than one category when confused, and the potential was computed as the proportion of responses each category got. Human accuracy at scene recognition was 90.4, 89.8, 86.8 and 85.3% for the different resolutions, as compared to the machine accuracy of 81.8%. Note that human performance is not 100% even with full resolution images because the scene categories are semantically ambiguous. Humans clearly outperform the machine at scene recognition, but the question of interest is whether this will translate to improved performance for scene understanding.

Similar to the class-class experiment, to obtain scene-class co-occurrence statistics, subjects were asked which object category is more likely to be present in the scene. We “show” the scene either by naming its category (no visual information), or by showing them the average image for that scene category³. Examples are shown in Figure 8. The normalized co-occurrence matrix is then used as the pairwise potential.

5.6 Ground truth Potentials

In addition to human potentials (which provide a lower-bound), we are also interested in establishing an upper-bound on the effect each subtask can have on segmentation performance by introducing ground truth (GT) potentials into the model. We formed each potential using the dataset annotations. For segments and super-segments we simply set the value of the potential to be 1 for the segment GT label and 0 otherwise, similarly for scene and class unary potentials. For object detection, we used the GT boxes as the candidates and set their detection scores to 1. For

3. When asked to look at the average images and recognize the scene category, subjects were 80% accurate.

the shape prior, we use a binary mask that indicates which pixels inside the GT object bounding box have the object’s label.

Note that in theory, some other settings of the variables in the model might produce better results than using ground truth. Therefore, using the ground truth information for each sub-task might not result in a strict upper-bound.

6 EXPERIMENTS WITH CRFs

We now describe the results of inserting the human potentials in the CRF model. We also investigated how plugging in GT potentials or discarding certain tasks all together affects performance on the MSRC dataset. For meaningful comparisons, CRF learning and inference is performed every time a potential is replaced, be it with (i) **Human** or (ii) **Machine** or (iii) **GT** or (iv) **Remove**.

A summary of the results for the four different settings is shown in Figure 9. Note that in each experiment only a *single* machine potential was replaced, which is indicated in the x axis of the plots. Missing bars for the *remove* setting indicate that removing the corresponding potential would result in the CRF being disconnected, and hence that experiment was not performed. GT is not meaningful for pairwise potentials. The average over all categories is shown on the y axis.

There are several interesting trends. Having GT information for class presence (i.e. knowing which objects are present in the image) clearly helps scene recognition (Figure 9(b)-last column), but also gives a noticeable boost to object detection (Figure 9(c)-last column) and segmentation (Figure 9(a)-last column). This argues in favor of informative classifiers for class presence, which were not used in the current model [7], but is, e.g., done in [21]. Class-class co-occurrence potential and the scene-class potential have negligible impact on the performance of all three tasks (5th and 7th columns in Figures 9(a), 9(b), and 9(c)). The choice of the scene classifier has little impact on the segmentation but influences detection accuracy (3rd column in Figures 9(a) and 9(c)). We find that human object detection boosts performance, which is not surprising (8th column in Figures 9(a), 9(b), and 9(c)). GT shape also improves segmentation performance (Figure 9(a)-4th column), but as discussed earlier, we find that humans are unable to instantiate this potential using the UCM segment boundaries. This makes it unclear what the realizable potential of shape is for the MSRC dataset.

One human potential that does improve performance is the unitary segment potential (1st column in Figures 9(a), 9(b), 9(c)). This is quite striking since human labeling accuracy of segments was substantially worse than machine’s (72.2% obtained by looking only at the segment pixels vs. 77.4% obtained

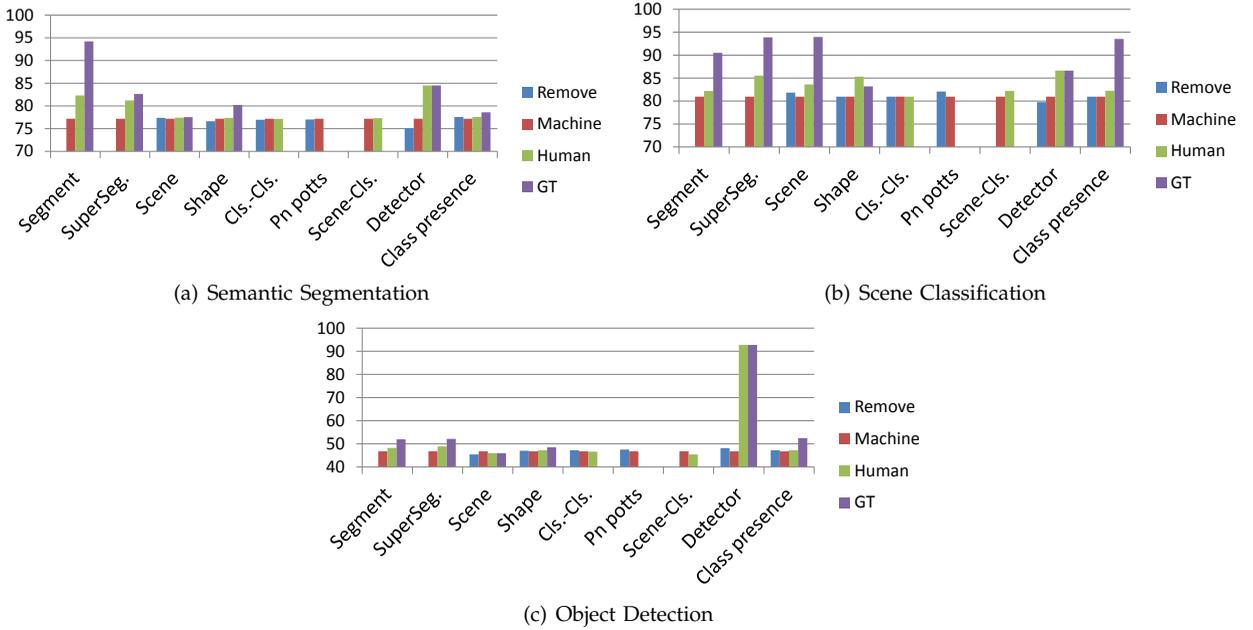


Fig. 9: Impact of each component on machine scene understanding. Here we show the semantic segmentation, object detection, and scene recognition accuracies when a single component of the model is changed (removed, implemented by a machine (default), replaced by a human or replaced with ground truth). The evaluation measures are average per-class recall, Average Precision (AP), and average recall for segmentation, object detection, and scene recognition respectively.

by the classifier, refer to Table 3), but incorporating the potential in the model significantly boosts performance (e.g., from 77.2% to 82.3% for the segmentation task). Note that for these experiments we use 200x200 windows for TextonBoost since it corresponds to the common practice in the literature, and allows us to match the set up in [7].

Intrigued by this, we performed a detailed analysis to identify properties of the human potential that are leading to this boost in performance. Resultant insights provided us concrete guidance to improve machine potentials and hence significantly better accuracies. This study is performed for the segmentation task.

6.1 Analysis of segments in MSRC

We now describe the various hypotheses we explored including unsuccessful and successful ones to explain the boost in the segmentation task provided by human segment potentials.

Scale: We noticed that the machine did not have access to the scale of the segments while humans did. So we added a feature that captured the size of a segment relative to the image and re-trained the unary machine potentials. The resulting segmentation accuracy of the CRF was 75.2%, unfortunately worse than the original accuracy at 77.2%.

Over-fitting: The machine segment unaries are trained on the same images as the CRF parameters, potentially leading to over-fitting. Humans obviously do

not suffer from such biases. To alleviate any over-fitting in the machine model, we divided the training data into 10 partitions. We trained the machine unaries on 9 parts, and evaluated them on the 10th part, repeating this 10 times. This gives us machine unaries on the entire training set, which can be used to train the CRF parameters. While the machine unaries may not be exactly calibrated, since the training splits are different by a small fraction of the images, we do not expect this to be a significant issue. The resultant accuracy was 76.5%, again, not an improvement.

Ranking of the correct label: It is clear that the highest ranked label of the human potential is wrong more often than the highest ranked label of the machine potential (hence the lower accuracy of the former outside the model). But we wondered if perhaps even when wrong, the human potential gave a high enough score to the correct label making it revivable when used in the CRF, while the machine was more “blatantly” wrong. We found that among the misclassified segments, the rank of the correct label using human potentials was 4.59 – not so different from 6.19 (out of 21) by the machine. Therefore, the rank of the correct label in human responses cannot explain the boost.

Uniform potentials for small segments: Recall that we did not have human subjects label the segments smaller than 500 pixels and assigned a uniform potential to those segments. The machine on the other hand produced a potential for each segment. We suspected that ignoring the small (likely to be misclassified) segments may give the human potential an advantage

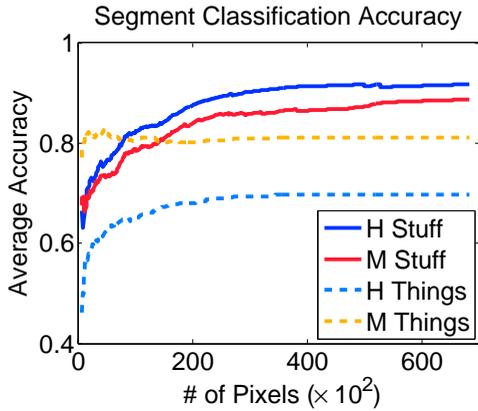


Fig. 10: Humans (H) and machines (M) have different performance for recognizing stuff and things segments. Humans are generally better at recognizing stuff, while machines are better at things recognition. Larger segments are generally easier to recognize. Note that in this experiment, Humans see only the segment pixels, while the machine classifier incorporates the information from the neighboring regions of the segments.

in the model. So we replaced the machine potentials for small segments with a uniform distribution over the categories. The average accuracy unfortunately dropped to 76.5%. As a follow-up, we also weighted the machine potentials by the size of the corresponding segment. The segmentation accuracy was still 77.1%, similar to the original 77.2%.

Regressing to human potentials: We then attempted to directly regress from the machine potential as well as the segment features (TextonBoost, LBP, SIFT, ColorSIFT, location and scale) to the human potential, with the hope that if for each segment, we can predict the human potential, we may be able to reproduce the high performance. We used Gaussian Process regression with an RBF kernel. The average accuracy in both cases was lower: 75.6% and 76.5%. We also replicated the sparsity of human potentials in the machine potentials, but this did not improve performance by much (77.3%).

Complementarity: To get a deeper understanding as to why human segment potentials significantly increase performance when used in the model, we performed a variety of additional hybrid CRF experiments. These included having human or machine potentials for segments or super-segments or both, with or without the P^n potential in the model. The results are shown in Table 2. The last two rows correspond to the case where both human and machine segment potentials are used together at the same level. In this case, using a P^n potential or not has little impact on the accuracy. But when the human and machine potentials are placed at different levels (recall that we have segments and super-segments at two different levels as shown in Fig. 1) in the model (rows 3 and 4), not having a P^n potential (and thus loos-

	P^n	without P^n
Human Seg., Human Super-seg.	78.9	77.2
Machine Seg., Machine Super-seg.	77.2	77.0
Human Seg., Machine Super-seg.	82.3	75.3
Machine Seg., Human Super-seg.	81.2	78.2
Human Seg. + Machine Seg., Machine Super-seg.	80.9	81.3
Human Seg. + Machine Seg., Human Super-seg.	82.3	82.8

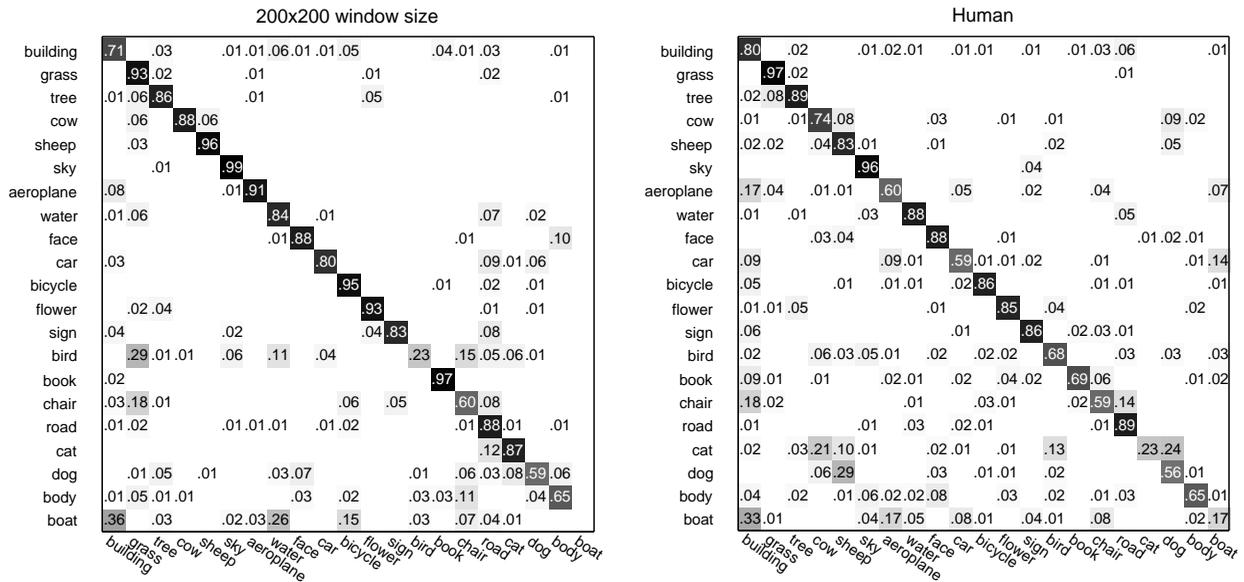
TABLE 2: Human and machine segment potentials are complementary. The last two rows correspond to the case where both human and machine segment potentials are used together at the same level. In this case, using a P^n potential or not has little impact on the accuracy. But when the human and machine potentials are placed at different levels in the model (rows 3 and 4), not having a P^n potential (and thus losing connection between the two levels) significantly hurts performance. This indicates that even though human potentials are not significantly more accurate than machine potentials, when both human and machine potentials interact, there is a significant boost in performance, demonstrating the complimentary nature of the two.

ing connection between the two levels) significantly hurts performance. This indicates that even though human potentials are not more accurate than machine potentials (obtained from a classifier that uses 200x200 windows), when both human and machine potentials interact, there is a significant boost in performance, demonstrating the complementary nature of the two.

Therefore, we hypothesized that the types of mistakes that the machine and humans make may be different. Our initial analysis showed that humans are generally better at detecting stuff while machine is better recognizing things (Figure 10).

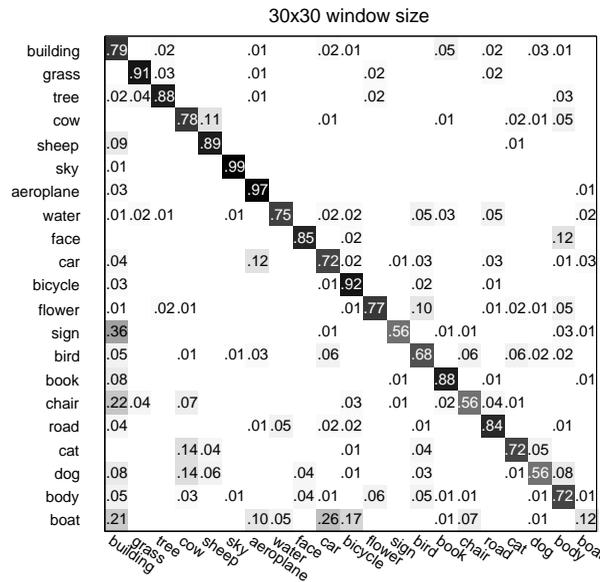
Additionally, we qualitatively analyzed the confusion matrices for both (Figure 11). We noticed that the machine with 200x200 windows confuses categories that spatially surround each other *e.g.*, bird and grass or car and road. This was also observed in [35]. This is understandable since a large window surrounding a pixel is used to generate its feature descriptor. On the other hand, human mistakes are between visually similar categories *e.g.*, car and boat.

We also analyzed the correlation of the mistakes within a super-segment and noticed that mistakes made within a super-segment are consistent for machines but variable for humans. Specifically, on average machine assigns different labels to 4.9% of segments, while humans assign different labels to 12% of the segments within a super-segment. This indicates that machine mistakes are structured and correlated so it may be harder for other components in the CRF to recover from these mistakes. Additionally, we analyzed the weights of the CRF to see what components become more effective when we use human segment potentials. We observed that the weight for the segment classification potential is lower for the case that we use human segment potentials. However, the weight for supersegment classification and P^n potential (which connects the supersegments



(a) Machine classification with 200x200 windows

(b) Human classification



(c) Machine classification with 30x30 windows

Fig. 11: The confusion matrices for segment classification are shown. For machine with large window size (a), there is high confusion between classes appearing in the surrounding area of each other, for example, bird-grass, car-road, etc. The types of mistakes are different for humans (b). They confused objects that look similar, for instance, there is confusion between cat-cow or boat-aeroplane. When we reduce the window size for machine (c), the mistakes become more similar to the human mistakes. Combining the small-window machine potentials with the large-window machine potentials results in a significant improvement in segmentation accuracy.

	200×200	10×10	20×20	30×30	40×40	Humans
in isolation	77.2	66.3	70.8	75.6	77.2	72.2
inside CRF	77.2	77.9	78.5	79.6	79.6	82.3

TABLE 3: Segmentation accuracy obtained by humans and by resizing TextonBoost window size outside the model (in isolation) and inside the CRF model. Note that, in the first row, the accuracy corresponds to segments larger than 500 pixels, while the second row shows the result for all segments.

to segments) increases significantly. This shows that in cases that the errors are decorrelated within a super-segment, smoothness terms become more effective.

One factor in the machine classifier that controls the correlation of classification mistakes of nearby segments is the TextonBoost window size. Our hypothesis is that as the window size decreases, the errors become less correlated, hence, we should observe more similar pattern of mistakes to human mistakes. Therefore, we plugged segment unaries that are computed based on small windows (as explained in Section 5.1) into the model. The average accuracy using window sizes of 10, 20, 30 and 40 are shown in the second row of Table 3. The accuracy outside the

model is shown in the first row of Table 3. Note that segment unaries computed based on 30x30 windows provide 2.4% improvement over [7]⁴. Notice that the improvement provided by the entire CRF model over the original machine unaries *alone* was 3% (from 74.2% to 77.2%). While a fairly straightforward change in the training of machine unaries leads to this improvement in performance, we note that the insight to do so was provided by our use of humans to “debug” the model.

6.2 Analysis of Segment Classification in PASCAL

PASCAL VOC dataset is labeled only with 20 “things”, making it uninteresting for scene understanding. Hence, we used our new dataset [10] that augments PASCAL VOC 2010 with annotations for 400+ additional classes. To perform the experiments in this section, we used 14 additional classes (shown in Table 4) that appear frequently at the immediate surrounding of the original 20 object classes. We used ~650 images for training the machine classifier and tested the learned classifier on ~200 random images from the validation set. We used the same MTurk setup as Figure 2 to carry out the human experiments on these 200 images.

Additional PASCAL classes						
sky	grass	ground	road	building	tree	water
mountain	wall	floor	railroad	keyboard	door	ceiling

TABLE 4: Additional categories used for segment classification in PASCAL VOC.

For classification, we used the method of [42]. The only difference is that we use patches of different sizes as the input to the classifier instead of their CPMC segments. We use square patches to be consistent with TextonBoost that was used for MSRC. The reason that we did not use TextonBoost for this experiment is that it does not scale well to larger number of images and categories. The patches are centered at the center of superpixels so there is a patch associated to each superpixel. We used 30x30 and 100x100 patches for our experiments. The confusion matrices are shown in the supplementary document due to space limitation.

A similar pattern of confusion exists in the PASCAL dataset. For example, for humans, there is confusion between aeroplane and semantically similar categories such as car and train, while the machine confuses aeroplane with sky or road that appear at immediate surroundings of aeroplanes and the confusion with car or train is negligible. Therefore, the types of machine and human mistakes are different for PASCAL dataset as well. Also, similar to the MSRC case, as we decrease the patch size, the machine error becomes more similar to the human

4. Adding a new unary potential simply by incorporating a different set of features and kernels than Textonboost (such as color, SIFT and self-similarity with intersection kernel) provides only a small boost at best (77.9%).

case. For instance, there is 31% confusion between boat and building when we use 100x100 patches, but when we reduce the patch size, the confusion becomes 51% (closer to 62% for humans). To quantify this, we computed the symmetric Kullback-Leibler (KL) distance between the corresponding rows of the confusion matrices and summed over the distances between the rows. The distance between Human and Machine 100x100 is 294.35 while it is 256.25 between Human and Machine 30x30. We see that the Machine 30x30 makes more human-like mistakes than Machine 100x100. Similarly, for MSRC, the distance between human and machine with large window was larger than human and machine with small window (151.2 vs. 63.8).

The accuracy for Machine with 100x100 patches is 28.1% and for human is 50.7%. To show that human and machine mistakes are complementary, we used an oracle that picked the segment label provided by humans if machine made a mistake. The pixel-wise accuracy of this approach is 56.5%, which is higher than both human and machine accuracies and shows the complementary nature of the mistakes. Designing machine potentials that make human-like mistakes may improve machine performance, as we demonstrated on the MSRC dataset in Section 6.1. Note that these experiments are independent of the system in [7].

We also performed an experiment to see if similar trends exist for the case that just like humans, machine also has access only to segment pixels (as opposed to a square patch that might include pixels outside the segment). We trained the machine classifier (similar as above) with all training images in [10] and evaluated the classifier on the same 200 images we used above. The average accuracy we obtained was 51.7%. As mentioned above, the accuracy for humans is 50.7%. Again, we used an oracle to choose the label chosen by humans whenever machine makes a mistake. The performance increased to 71.8%, which indicates the human and machine mistakes are complementary in this case too and the same trends as the above experiments exist.

7 ANALYZING THE PIPELINE

In this section, we analyze the scene understanding CRF model using the MSRC dataset. First, we investigate whether the components used in the model are beneficial for humans. Second, we estimate the potential that the model as a whole holds for all the three tasks. Note that the components that [7] use, e.g., segment classifier, object detector, etc., are common among most scene understanding approaches (e.g., [32], [43], [44], etc). So it is conceivable that our conclusions may generalize to other scene understanding approaches. However, we only analyze [7], and have not tested the generality of our conclusions to other pipelines.

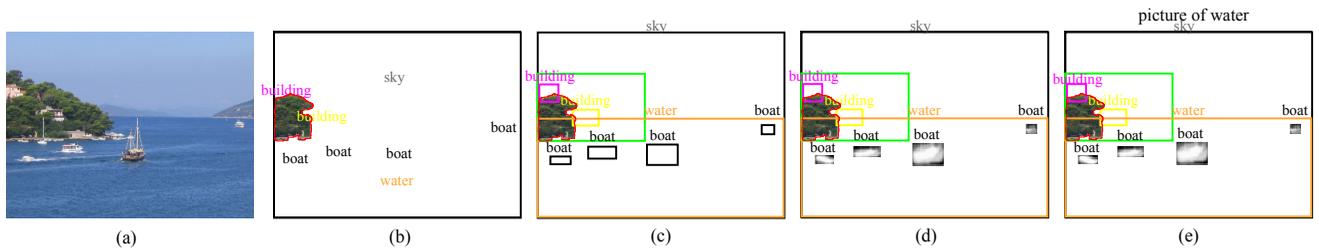


Fig. 12: (b):(e) Human-study interfaces for (a) with increasingly more contextual information.

7.1 Contextual Image Labeling

To study if humans benefit from contextual information provided by the model in order to recognize local regions, we design an experiment where, in addition to the segment pixels, we show subjects progressively more information from the model: presence / absence of each class, object bounding boxes, shape prior masks and scene type. We selected 50 random segments of any size from each category and asked the subjects to classify each segment. Figure 12 shows an example for each of the interfaces used for this study. From left to right (b – e), the contextual information available to the subjects is increased.

The results are shown in Table 5. We see that human performance in the task of image labeling significantly increases when presented with the same contextual information used by machine models. We also show accuracies of the machine model when given access to precisely the same information as shown to humans. We find that humans can effectively leverage the available information. Even though human performance is worse than machines when viewing segments in isolation, they outperform the machine when given access to the contextual information modeled by the CRF. This is especially true for “things”. Access to context seem to confuse human responses for “stuff”. Investigating this further is part of future work.

7.2 Journey to Perfection

To analyze if the model has the potential to reach perfection, we conduct experiments using different combinations of machine, human, and ground truth potentials. Figure 13 provides a journey on the segmentation task from the machine’s current 77% performance to perfection. We find that incorporating human (H) segment (S) potentials improves performance by 5-6%. Incorporating ground truth (GT) detection (Det) provides an improvement of 4-7%. Adding in GT for remaining tasks (except super-segments (SS)) further improves performance by 2-5%. These combined bring us to 92%. GT segments plugged into the model perform at 94.2%, which outside the model are at 94.5% (the upper-bound on the performance of the model since it makes segment-level decisions). This shows that as far as segmentation goes the model itself is sufficient and all the required tasks are being modeled. This analysis also provides concrete guidance for future research efforts: designing representations complementary to the ones

used in the model, perhaps by mimicking human responses, has potential for significant improvement. And of course, improving on all the tasks would lead to more effective scene understanding. Note that, holistically accurate systems do not require extremely high performance on each sub-task to achieve high performance.

Figure 14 shows the effect of each component on the object detection task. The machine’s average precision is 46.8. Including the ground truth information for class presence (CP) improves the average precision to 52.5 (5.7 improvement in AP). Incorporating human segment (S) and super-segment (SS) potentials provides an additional improvement of 1.1 AP, which leads to 53.6 AP. Ground truth shape information also provides improvement for the object detection task, and increases the AP to 54.3. Including the ground truth scene has negligible impact on performance. If we replace the human segment and super-segment potentials by their ground truth counterparts, the average precision decreases to 54.0. Hence, ground truth class-presence (knowing whether a certain category exists in the image or not similar to the image classification task in PASCAL) is the single component that provides the biggest boost in performance. Obviously, using ground truth information for object detection has a significant effect on the performance, where it improves the AP to 93.9. Hence, given the scope of this model, the burden of improving the detection performance from 54.3 to 93.9 lies on the detector itself. Enhancing the model with additional cues such as a rough 3D layout of the scene, etc. that directly influence likely locations and scales of detections may be important to aid the detector.

The journey for scene recognition is shown in Figure 15. The machine performance for scene recognition is 81.0%. Using ground truth shape potential improves the performance by 2.2%. Using ground truth detection (Det), segment (S) and super-segment (SS) potentials instead provides 9.3% boost in accuracy. Adding ground truth shape to this combination does not change the accuracy. Using ground truth class-presence single-handedly takes the performance to 93.5% (an improvement of 12.5%, while ground truth scene information is at 94.0%). This is because the scene categories in this dataset as defined in Table 1 are object-category centric, and hence knowing whether a certain object-category is present in an image or not provides strong cues about the scene

	build.	grass	tree	cow	sheep	sky	aeropl.	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	avg.	avg. stuff	avg. things
Humans																								
Segm. (S)	48	82	38	20	14	64	18	60	50	22	12	58	44	8	20	16	36	0	6	22	6	30.7	54.7	21.1
(b) S+label	78	84	54	86	82	58	64	62	72	50	72	90	92	88	90	90	52	92	82	58	70	74.6	64.7	78.5
(c) S+box (B)	58	72	38	100	98	44	100	54	82	96	94	100	96	96	98	96	46	100	98	92	98	83.6	52.0	96.3
(d) S+B+msk	58	72	44	100	100	42	100	54	98	98	94	100	96	94	98	98	36	100	98	96	100	84.6	51.0	98.0
(e) Full info.	68	78	48	98	100	50	100	66	100	98	94	100	96	94	98	98	46	100	98	90	100	86.7	59.3	97.6
Machine model																								
Segm. (S)	82	86	93	74	94	96	84	88	96	70	90	88	80	6	97	30	92	95	51	34	0	72.7	89.5	65.9
S+label	83	87	93	76	92	97	87	92	96	73	100	98	80	42	97	54	93	96	70	33	0	78.2	90.8	72.9
S+box (B)	83	87	93	86	96	97	90	92	96	82	100	100	87	44	100	61	93	96	71	35	55	82.9	90.8	79.9
S+B+msk	84	87	93	86	98	97	90	92	97	82	100	100	87	44	100	61	93	96	71	42	75	84.3	91.0	81.9
Full info.	84	87	93	86	98	97	90	92	96	82	100	100	87	44	100	61	93	96	71	42	75	84.4	91.0	81.9

TABLE 5: Human segmentation accuracies with increasing information from the model. The letters in the first column of the Human experiment correspond to the letters in Fig. 12.

category.

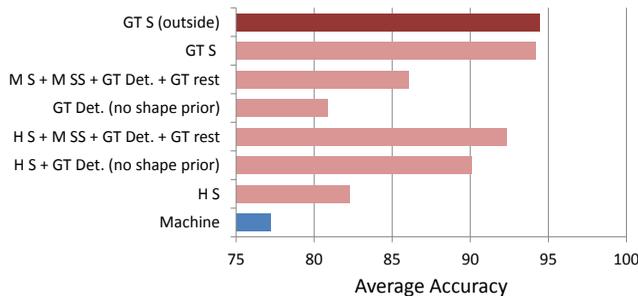


Fig. 13: Journey to perfection for segmentation.

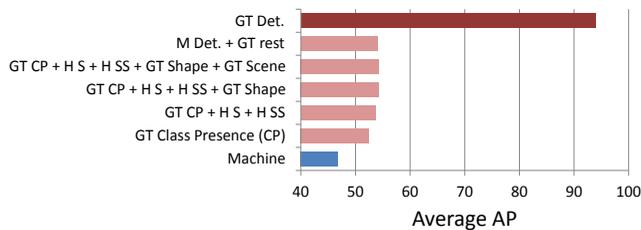


Fig. 14: Journey to perfection for object detection.

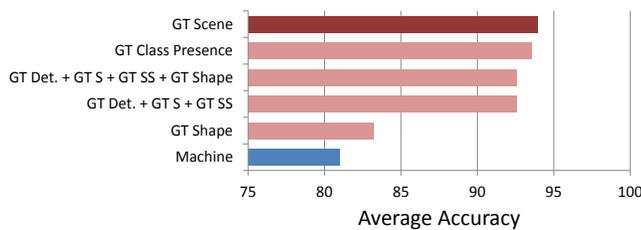


Fig. 15: Journey to perfection for scene classification.

8 CONCLUSION

Researchers have developed sophisticated machinery for scene understanding. Insights into which aspects of these models are crucial, especially for further improving state-of-the-art performance is valuable. We gather these insights by analyzing a CRF model for scene understanding.

Our analysis hinges on the use of human subjects to produce the different potentials in the model. Comparing performance of various human-machine hybrid models allows us to identify the components of the model that still have “head room” for improving

performance. One of our findings was that human responses to local segments in isolation, while being less accurate than machines’, provide complementary information that the CRF model can effectively exploit. We showed that a similar pattern of mistakes happens for the more difficult PASCAL dataset. We explored various avenues to precisely characterize this complementary nature. We also investigated different shape priors for the model, and it turned out that human subjects can not decipher the object shape from super-pixel boundaries any better than machines. In addition, we showed that humans can effectively leverage the contextual information incorporated in the machine model.

We expect even more insightful findings if this model is studied on larger and more challenging datasets, which is part of future work.

Acknowledgments This work was supported in part by NSF IIS-1115719. The first author was partly supported by ONR grant N00014-12-1-0883. We would also like to thank Xianjie Chen for his help with some of the experiments.

REFERENCES

- [1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models.” *PAMI*, vol. 32, no. 9, 2010.
- [2] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” in *CVPR*, 2010.
- [3] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, “Objects in context,” in *ICCV*, 2007.
- [4] Y. Yang and D. Ramanan, “Articulated pose estimation using flexible mixtures of parts,” in *CVPR*, 2011.
- [5] A. Torralba, “How many pixels make an image?” *Visual Neuroscience*, 2009.
- [6] S. Gould, T. Gao, and D. Koller, “Region-based segmentation and object detection,” in *NIPS*, 2009.
- [7] J. Yao, S. Fidler, and R. Urtasun, “Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation,” in *CVPR*, 2012.
- [8] G. Heitz, S. Gould, A. Saxena, and D. Koller, “Cascaded classification models: Combining models for holistic scene understanding,” in *NIPS*, 2008.
- [9] C. Li, A. Kowdle, A. Saxena, and T. Chen, “Towards holistic scene understanding: Feedback enabled cascaded classification models,” in *NIPS*, 2010.
- [10] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, “The role of context for object detection and semantic segmentation in the wild,” in *CVPR*, 2014.

- [11] H. Barrow and J. Tenenbaum, "Recovering intrinsic scene characteristics from images," *In Comp. Vision Systems*, 1978.
- [12] D. Hoiem, A. A. Efros, , and M. Hebert, "Closing the loop on scene interpretation," in *CVPR*, 2008.
- [13] V. Lempitsky, P. Kohli, C. Rother, and B. Sharp, "Image segmentation with a bounding box prior," in *ICCV*, 2009.
- [14] T. Brox, L. Bourdev, S. Maji, and J. Malik, "Object segmentation by alignment of poselet activations to image contours," in *CVPR*, 2011.
- [15] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik, "Recognition using regions," in *CVPR*, 2009.
- [16] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *NIPS*, 2005, pp. 1401-1408.
- [17] E. Sudderth, A. Torralba, W. T. Freeman, and A. Wilsky, "Learning hierarchical models of scenes, objects, and parts," in *ICCV*, 2005.
- [18] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *ICCV*, 2005.
- [19] L. Ladicky, P. Sturges, K. Alahari, C. Russell, and P. H. Torr, "What, where and how many? combining object detectors and crfs," in *ECCV*, 2010.
- [20] C. Wojek and B. Schiele, "A dynamic conditional random field model for joint labeling of object and scene classes," in *ECCV*, vol. 4, 2008, pp. 733-747.
- [21] G. Cardinal, X. Boix, J. van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez, "Harmony potentials for joint classification and segmentation," in *CVPR*, 2010.
- [22] J. Rivest and P. Cabanagh, "Localizing contours defined by more than one attribute," *Vision Research*, 1996.
- [23] C. C. Fowlkes, "Measuring the ecological validity of grouping and figure-ground cues," *Thesis*, 2005.
- [24] L. Fei-Fei, R. VanRullen, C. Koch, and P. Perona, "Rapid natural scene categorization in the near absence of attention," *PNAS*, 2002.
- [25] T. Bachmann, "Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity?" *Europ. J. of Cogn. Psych.*, 1991.
- [26] A. Oliva and P. G. Schyns, "Diagnostic colors mediate scene recognition," *Cognitive Psychology*, 2000.
- [27] D. Parikh and C. L. Zitnick, "The role of features, algorithms and data in visual recognition," in *CVPR*, 2010.
- [28] D. Parikh and C. Zitnick, "Finding the weakest link in person detectors," in *CVPR*, 2011.
- [29] D. Parikh, "Recognizing jumbled images: The role of local and global information in image classification," in *CVPR*, 2011.
- [30] C. L. Zitnick and D. Parikh, "The role of image understanding in contour detection," in *CVPR*, 2012.
- [31] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh, "Analyzing semantic segmentation using human-machine hybrid crfs," in *CVPR*, 2013.
- [32] L. Ladicky, C. Russell, P. H. S. Torr, and P. Kohli, "Associative hierarchical crfs for object class image segmentation," in *ICCV*, 2009.
- [33] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Distributed message passing for large scale graphical models," in *CVPR*, 2011.
- [34] T. Hazan and R. Urtasun, "A primal-dual message-passing algorithm for approximated large scale structured prediction," in *NIPS*, 2010.
- [35] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textronboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context," *IJCV*, vol. 81, no. 1, 2007.
- [36] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," in *BMVC*, 2007.
- [37] J. Shotton, M. Johnson, and R. Cipolla, "Semantic textron forests for image categorization and segmentation," in *CVPR*, 2008.
- [38] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," in *PAMI*, 2011.
- [39] P. Kohli, M. P. Kumar, and P. H. S. Torr, " p^3 and beyond: Solving energies with higher order cliques," in *CVPR*, 2007.
- [40] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, p. 462467, 1968.
- [41] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *ICCV*, 2011.
- [42] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *ECCV*, 2012.
- [43] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. S. Gual, and J. Gonzalez, "Harmony potentials - fusing global and local scale for semantic image segmentation." *IJCV*, 2012.
- [44] M. Sun, B. soo Kim, P. Kohli, and S. Savarese, "Relating things and stuff via objectproperty interactions," *PAMI*, 2014.



Roozbeh Mottaghi is currently a research scientist at the Allen Institute for Artificial Intelligence. He obtained his PhD in Computer Science from UCLA in 2013. His research interests are scene understanding and object detection and segmentation.



Sanja Fidler is an Assistant Professor at the Computer Science department of University of Toronto. Before that, she was a Research Assistant Professor at TTI-Chicago. She completed her PhD in computer science at University of Ljubljana, and was a post-doctoral fellow at University of Toronto. Her research interests are object detection and segmentation, 3D scene understanding, and combining language and vision.



Alan Yuille received the BA degree in mathematics from the University of Cambridge in 1976. His PhD on theoretical physics, supervised by Prof. S.W. Hawking, was approved in 1981. He was a research scientist in the Artificial Intelligence Laboratory at MIT and the Division of Applied Sciences at Harvard University from 1982 to 1988. He served as an assistant and associate professor at Harvard until 1996. He was a senior research scientist at the Smith-Kettlewell Eye Research Institute from 1996 to 2002. He joined the University of California, Los Angeles, as a full professor with a joint appointment in statistics and psychology in 2002. He obtained a joint appointment in computer science in 2007. His research interests include computational models of vision, mathematical models of cognition, and artificial intelligence and neural networks.



Raquel Urtasun is an Assistant Professor at the Computer Science department of University of Toronto. Previously, she was an Assistant Professor at TTI-Chicago and was a postdoctoral research scientist at UC Berkeley and ICSI and a postdoctoral associate at the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT. Raquel Urtasun completed her PhD at the Computer Vision Laboratory at EPFL, Switzerland. Her major interests are statistical machine learning and computer vision.



Devi Parikh is an Assistant Professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech (VT), where she leads the Computer Vision Lab. Prior to this, she was a Research Assistant Professor at Toyota Technological Institute at Chicago (TTIC). She received her Ph.D. from the Electrical and Computer Engineering department at Carnegie Mellon University in 2009. She received her B.S. in Electrical and Computer Engineering from Rowan University in 2005. Her research interests include computer vision, pattern recognition and AI in general and visual recognition problems in particular.