Analyzing Semantic Segmentation Using Hybrid Human-Machine CRFs

Roozbeh Mottaghi
UCLA
roozbehm@cs.ucla.edu

Sanja Fidler, Jian Yao, Raquel Urtasun
TTI Chicago
{fidler,yaojian,rurtasun}@ttic.edu

Devi Parikh
Virginia Tech
parikh@vt.edu

Abstract

Recent trends in semantic image segmentation have pushed for holistic scene understanding models that jointly reason about various tasks such as object detection, scene recognition, shape analysis, contextual reasoning. In this work, we are interested in understanding the roles of these different tasks in aiding semantic segmentation. Towards this goal, we "plug-in" human subjects for each of the various components in a state-of-the-art conditional random field model (CRF) on the MSRC dataset. Comparisons among various hybrid human-machine CRFs give us indications of how much "head room" there is to improve segmentation by focusing research efforts on each of the tasks. One of the interesting findings from our slew of studies was that human classification of isolated super-pixels, while being worse than current machine classifiers, provides a significant boost in performance when plugged into the CRF! Fascinated by this finding, we conducted in depth analysis of the human generated potentials. This inspired a new machine potential which significantly improves state-of-the-art performance on the MRSC dataset.

1. Introduction

We consider the problem of semantic image segmentation. Clearly, other image understanding tasks like object detection [10], scene recognition [38], contextual reasoning among objects [29], and pose estimation [39] can aid semantic segmentation. For example, knowing that the image is a street scene influences where we expect to find people. 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 [35]. In fact, different and functionally complementary regions in the brain are known to co-operate to perform scene understanding [28].

Recent works [12, 40, 16, 23], have thus pushed on *holistic* scene understanding models for among other things, improved semantic segmentation. 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 performances in a variety of benchmarks.

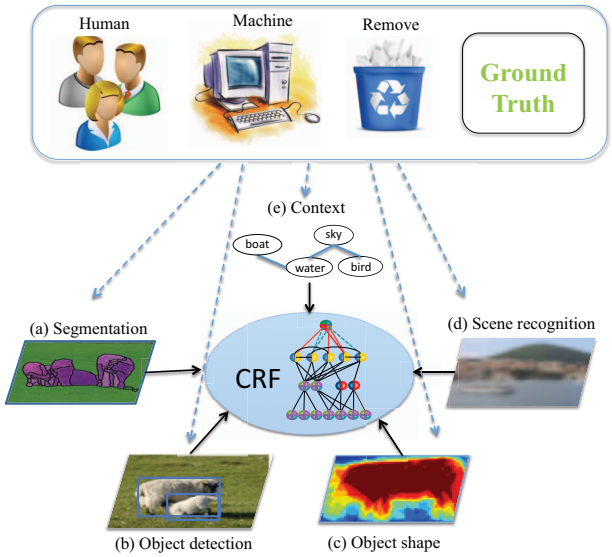


Figure 1. A holistic scene understanding approach to semantic segmentation consists of a conditional random field (CRF) model that jointly reasons about: (a) classification of local patches (segmentation), (b) object detection, (c) shape analysis, (d) scene recognition and (e) contextual reasoning. In this paper we analyze the relative importance of each of these components by building an array of hybrid human-machine CRFs where each component is performed by a machine (default), or replaced by human subjects or ground truth, or is removed all together (top).

In this paper, we aim to determine the relative importance of the different recognition tasks in aiding semantic segmentation. Our goal is to discover which of the tasks if improved, can boost segmentation performance significantly. In other words, to what degree can we expect to improve segmentation 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 state-of-art methods for semantic segmentation.

We analyze the recent and most comprehensive holistic scene understanding model of Yao *et al*. [40]. It is a conditional random field (CRF) that models the interplay between segmentation and a variety of components such as local super-pixel appearance, object detection, scene recognition, shape analysis, 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 one, and substitute a machine with a human for that task, keeping the rest of the model intact (Figure 1). The resultant improvement in segmentation performance, if any, will give us an indication of how much "head room" there is to improve segmentation 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[1]. 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 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 in fact less accurate than machine classification. However when plugged into the holistic 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 (e.g. complementary 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 the state-of-the-art on the MSRC dataset.

2. Related Work

Holistic Scene Understanding: The key motivation behind holistic scene understanding, going back to the seminal work of Barrow in the seventies [3], is that ambiguities in visual information can only be resolved when many visual processes are working collaboratively. A variety of holistic approaches have since been proposed. Many of these works incorporate the 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) [17, 16, 22, 5, 13]. There are fewer efforts on joint reasoning of the various recognition tasks. In [36], contextual information was incorporated into a CRF leading to improved object detection. A hierarchical generative model spanning parts, objects and scenes is learnt in [34]. Joint estimation of depth, scene type, and object locations is performed in [23]. Spatial contextual interactions between objects have also been modeled [19, 29]. Image segmentation and object detection are jointly modeled in [21, 37, 12] using a CRF. [6] 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 semantic segmentation.

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 [30] found that luminance, color, motion and texture cues for contour detections are integrated at a common site in the brain. Fowlkes [11] 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 semantic segmentation which involves identifying the semantic category of each pixel in the image.

Several works have studied high-level recognition tasks in humans. Fei-Fei *et al*. [9] show that humans can recognize scenes rapidly even while being distracted. Bachmann *et al*. [2] 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*. [35] show that humans can reliably detect objects in 32×32 images. In contrast, in this paper, we study human performance at tasks that closely mimic existing holistic computational models for semantic segmentation in order to identify bottlenecks, and better guide future research efforts.

Parikh *et al*. [27] recently applied human studies to identify the weakest links in existing models for the specific task of person detection. In contrast, 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 *holistic scene understanding* models for semantic segmentation. While similar at the level of exploiting human involvement, the problem, the model, the methodolo-

[1] 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 segmentation performance.

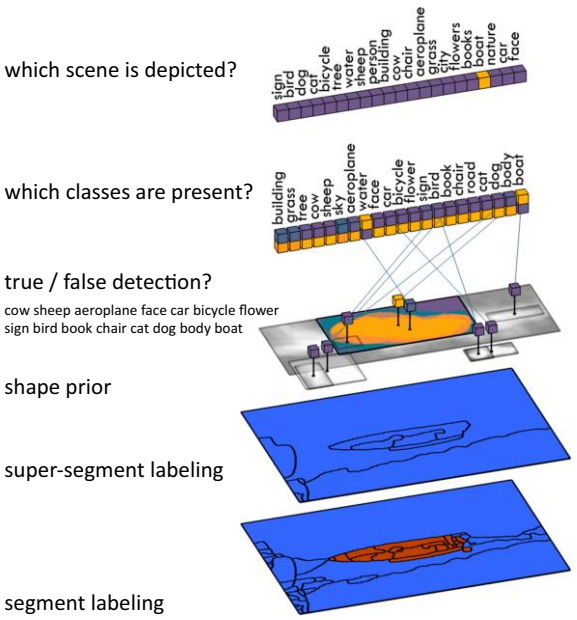


Figure 2. Overview of the holistic scene model of [40] that we analyze using human subjects. For clarity, not all connections in the model are shown here.

gies of the human studies and machine experiments, as well as the findings and insights are all novel.

3. CRF Model

We analyze the recently introduced CRF model of [40] which reasons jointly about a variety of scene components. While the model shares similarities with past work [20, 21, 6], we choose this model because it provides state-of-the-art performance in holistic scene understanding, and thus forms a great starting point to ask "which components need to be improved to push the state-of-the-art 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 [31] 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 2). We refer the reader to [40] for further technical details. The problem of holistic 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 [40], we will refer to them as segments and super-segments. The model also has binary variables indicating the correctness of candidate object detection bounding boxes. In addition, a multi-label 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, i.e., 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 that respect this prior 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 $x_i \in \{1, \cdots, C\}$ and $y_j \in \{1, \cdots, C\}$ 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 [10] 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 $s \in \{1, \ldots, C_l\}$ 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 [15]. 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 [33] semantic labeling benchmark, also used by [40]. The MSRC dataset is widely used, contains stuff (e.g., *sky, water*), things (i.e., shape-defined classes such as *cow, car*) and a diverse set of scenes, making it a good choice among existing datasets for our study[2]. We use the more precise groundtruth of MSRC pro-

[2]The PASCAL dataset is more challenging in terms of object ("things") detection and segmentation. However, a large portion of its images, especially "stuff", is unlabeled. The contextual interactions are also quite skewed [7] making it less interesting for holistic scene understanding. The SUN dataset [38] is prohibitively large for the scale of human studies involved in our work. The SIFT-flow dataset [24] is dominated by "stuff" with a small proportion of "things" pixels. Camvid [4] is limited to street scenes.

vided by Malisiewicz and Efros [25] and used in [40], as it offers a more accurate measure of performance. We use the same scene category and object detection annotations as in [40]. Figure 2 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 [21]) as well. We use the standard train/test split from [32] to train all machine potentials, described next.

5. Machine CRF Potentials

We now describe the machine potentials we employed. Our choices closely follow those made in [40].

Segments and super-segments: We utilize UCM [1] to create our segments and super-segments as it returns a small number of segments that tend to respect the true object boundaries well. 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 [33] in [20] to get pixel-wise potentials and average those within the segments and super-segments to get the unary potentials. Following [18], we connect the two levels via a pairwise P^n potential that encourages segments and super-segments to take the same label.

Class: 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-occurance statistics of pairs of classes. However, for efficiency reasons, instead of utilizing a fully connected graph, we use a tree-structure obtained via the Chow-Liu algorithm [8] on the class-class co-occurrence matrix.

Detection: Detection is incorporated in the model by generating a large set of candidate bounding boxes using the deformable part-based model [10]. 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.

Shape: 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, 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.

Scene and scene-class co-occurrence: We train a classifier [38] 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-occurance statistics of the training data, i.e., likelihood of each class being present for each scene type.

6. Human CRF Potentials

We now explain our human studies. Section 7 presents the results of feeding these human "potentials" into the machine model. 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 instructions. 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.

Segments and Super-segments: The study involves having human subjects classify segments into one of the semantic categories. Subjects were only shown pixels that belong to the segment. The segment was shown within a rectangle corresponding to the image around it, making its location and scale in the image evident. If confused, subjects were allowed to select multiple classes for each segment. See Figure 3. The machine classifier, TextonBoost [33] in particular, has access to a large neighborhood (200x200 pixels) around the segment. Clearly, it does not use information only from the pixels in the segment while classifying the segment. However, showing all the information that the machine uses to human subjects would lead to nearly 100% classification accuracy by the subjects, leaving us with little insights to gain. More importantly, a 200 x 200 window occupies nearly 60% of the image, resulting in humans potentially using holistic scene understanding while classifying the segments. This would contradict our goal of having humans perform individual tasks in isolation. Finally, a direct comparison between humans and machines is not of interest to us. We are interested in understanding the potential each component in the model holds. To this goal, the discrepancy in information shown to humans and machines is not a concern, as long as humans are not shown *more* information than the machine has access to. 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 class, or allowing a subject to select only one category per segment, etc.). The one shown in Figure 3 resulted in most consistent responses from subjects.

Our experiment involved having subjects label all segments and super-segments from the MSRC dataset containing more than 500 pixels. This resulted in 10976 segments

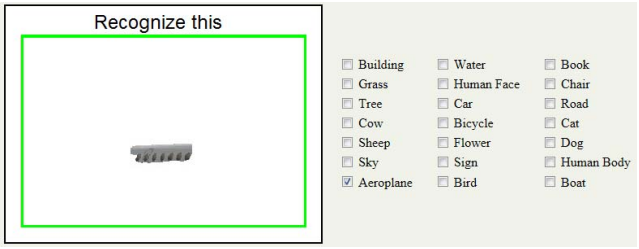


Figure 3. Segment labeling interface.

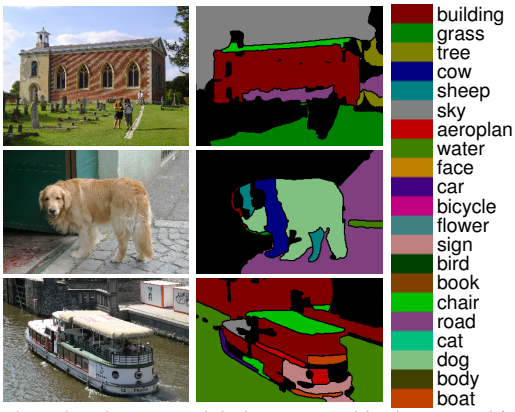


Figure 4. Isolated segment labels generated by human subjects

and 6770 super-segments. They cover 90.2% and 97.7% of all pixels in the dataset[3]. Figure 4 shows examples of segmentations obtained 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%, as compared to 77.4% for machines[4]. As expected, humans perform rather poorly when only local information is available. Accuracy for super-segments is 84.3% and 79.6% respectively. 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.

Class Unary: 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.

Class-Class Co-occurrence: 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 [40] on the class co-occurrence potentials to obtain the tree structure, where the edges connect highly co-occurring nodes. The structure of the human tree is quite similar to the tree obtained from the MSRC training set. Visualizations of the trees are available on author's webpage.

Object Detection: 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%.

Shape: 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. See Figure 5 [5]. Using the interface of [14], 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 corresponding accuracy for the detector-based shape prior snapped to the segments is 78.5%, not much worse than the human subjects. This shows that humans can not decipher the shape of an object from the UCM segment boundaries better than an automatic approach. Hence, it is unlikely that simply "puzzling together" UCM-like segments will improve shape analysis.

Scene Unary: We ask human subjects to classify an image into one of the 21 scene categories used in [40] (see Figure 2). Images were presented at varying resolutions (i.e. original resolution, smallest dimension rescaled to 32, 24 and 20 pixels). Subjects were allowed to select

[3]Covering 100% of the pixels in the dataset would involve labeling three times the number of segments, and the resources seemed better utilized in the other human studies.

[4]This accuracy is calculated only over segments larger than 500 pixels that were shown to humans. Machine accuracy over all segments is 74.2%.

[5]We showed subjects 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 (54% accuracy). With context, classification accuracy was 94.0%.

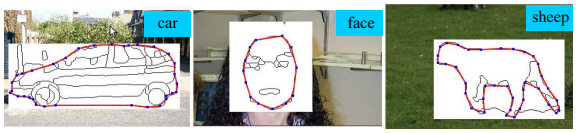


Figure 5. Shape mask labelling interface.

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 semantic segmentation performance.

Scene-Class Co-occurrence: Similar to the class-class experiment, 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 category. The normalized co-occurrence matrix is then used as the pairwise potential.

Ground-truth Potentials: In addition to human potentials (which provide a feasible point), we are also interested in establishing an upper-bound on the effect each subtask can have on segmentation performance. We do so 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 the shape prior, we use a binary mask that indicates which pixels inside the GT object bounding box have the object's label.

7. Experiments with Human-Machine 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 segmentation 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 6. Note that in each experiment only a *single* machine potential was replaced, which is indicated in the x axis of the plot. 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

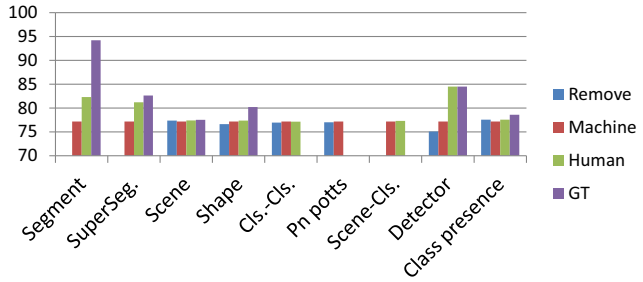


Figure 6. Impact of each component on semantic segmentation.

potentials. The average recall is shown on the y axis. Due to space considerations, we provide detailed class-wise accuracy tables in a separate document on the author's webpage.

Class presence, class-class co-occurrence, and the scene-class potentials have negligible impact on the performance of semantic segmentation. The choice of the scene classifier also has little impact on this dataset. We find that GT object detection boosts performance, which is not surprising. GT shape also improves performance, 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. This is quite striking since human labeling accuracy of segments was substantially worse than machine's (72.2% vs. 77.4%), but incorporating the potential in the model significantly boosts performance (from 77.2% to 82.3%). Intrigued by this, we performed 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 state-of-the-art accuracies. We now describe the various hypotheses we explored (including unsuccessful and successful ones).

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 resultant 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 10^{th} 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 – noticeably better than 6.19 (out of 21) by the machine.

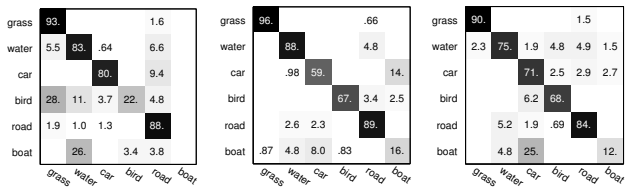
Uniform potentials for small segments: Recall that we did not have human subjects label the segments smaller than 500 pixels and assigned a uniform potential to those segments. The machine on the other hand produced a potential for each segment. We suspected that ignoring the small (likely to be misclassified) segments may give the human potential an advantage 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 77.1%, almost the same as 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 a Gaussian Process regressor with 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 CRF experiments with hybrid potentials. These included having human (H) or machine (M) potentials for segments (S) or super-segments (SS) or both, with or without the P^n potential in the model. The results are shown in Table 1. 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 more accurate than machine potentials, when both human and machine potentials interact, there is a significant boost in performance, demonstrating the complementary nature of the two.

	P^n	without P^n
H S, H SS	78.9	77.2
M S, M SS	77.2	77.0
H S, M SS	82.3	75.3
M S, H SS	81.2	78.2
H S+M S, M SS	80.9	81.3
H S+M S, H SS	82.3	82.8

Table 1. Human & machine segment potentials are complementary



(a) M: 200x200 windows (b) Human (c) M: 30x30 windows

Figure 7. (Sub) confusion matrices for isolated segment classification. M = machine.

So we hypothesized that the types of mistakes that the machine and humans make may be different. We qualitatively analyzed the confusion matrices for both. We noticed that the machine confuses categories that spatially surround each other e.g. bird and grass or water and boat (Figure 7(a)). This was also observed in [33] and is understandable because TextonBoost uses a large (200×200) window surrounding a pixel to generate its feature descriptor. On the other hand, human mistakes are between visually similar categories e.g. car and boat (Figure 7(b)).[6] Hence, we trained TextonBoost with smaller windows. The resultant confusion matrix was more similar to that of human subjects (Figure 7(c)). For the full confusion matrix refer to the author's webpage. We re-computed the segment unaries and plugged them into the model in addition to the original unaries that used large windows. The average accuracy we obtained by the model using window sizes of 10, 20, 30 and 40 were 77.9, 78.5, 79.6 and 79.6 (compare to 77.2%). This improvement of 2.4% over state-of-the-art is quite significant for this dataset[7]! Notice that the improvement provided by the entire CRF model over the original machine segment unaries *alone* was 3% (from 74.2% to 77.2%). While a fairly straightforward change in the training of machine unaries lead to this improvement in performance, we note that the insight to do so was provided by our use of humans to "debug" the state-of-the-art model.

[6] One consequence of this is that the 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. The consistent mistakes may be harder for other components in the CRF to fix.

[7] 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%).

Potential of the pipeline: Of course, in spite of MSRC being a well studied dataset, there is still room for improvement. GT labels for segments when plugged into the CRF provide an accuracy of 94% (and not 100% because decisions are made at the segment level which are not perfect). We find that not just the dataset, but even the particular model of Yao *et al*. [40] that we analyze in this paper has further potential. Plugging in human potentials for all the components gives us an accuracy of 89.5%. Our analysis reveals precisely which directions to pursue to reach this potential. We expect even more insightful findings if this model is studied on larger and more challenging datasets like the SUN dataset [38], which is part of future work.

8. Conclusion

Researchers have developed sophisticated machinery for semantic segmentation of images. 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 state-of-the-art CRF model for semantic segmentation on the MSRC dataset. Our analysis hinges on the use of human subjects to produce the different potentials in the model. Comparing performance of various hybrid human-machine models allows us to identify the components of the model that still have "head room" for improving segmentation 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 explored various avenues to precisely characterize this complementary nature, which resulted in a novel machine potential that significantly improves accuracy over the state-of-art.

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. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. In *PAMI*, 2011. 4

[2] 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. 2

[3] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. *In Comp. Vision Systems*, 1978. 2

[4] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008. 3

[5] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011. 2

[6] 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. 2, 3

[7] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 3

[8] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462467, 1968. 4

[9] L. Fei-Fei, R. VanRullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *PNAS*, 2002. 2

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010. 1, 3, 4

[11] C. C. Fowlkes. Measuring the ecological validity of grouping and figure-ground cues. *Thesis*, 2005. 2

[12] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*, 2009. 1, 2

[13] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, 2009. 2

[14] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 5

[15] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010. 3

[16] G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2008. 1, 2

[17] D. Hoiem, A. A. Efros, , and M. Hebert. Closing the loop on scene interpretation. In *CVPR*, 2008. 2

[18] P. Kohli, M. P. Kumar, and P. H. S. Torr. p^3 and beyond: Solving energies with higher order cliquess. In *CVPR*, 2007. 4

[19] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005. 2

[20] L. Ladicky, C. Russell, P. H. S. Torr, and P. Kohli. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 3, 4

[21] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, 2010. 2, 3, 4

[22] V. Lempitsky, P. Kohli, C. Rother, and B. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009. 2

[23] C. Li, A. Kowdle, A. Saxena, and T. Chen. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *NIPS*, 2010. 1, 2

[24] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: label transfer via dense scene alignment. In *CVPR*, 2009. 3

[25] T. Malisiewicz and A. A.Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007. 4

[26] A. Oliva and P. G. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 2000. 2

[27] D. Parikh and C. Zitnick. Finding the weakest link in person detectors. In *CVPR*, 2011. 2

[28] S. Park, T. Brady, M. Greene, and A. Oliva. Disentangling scene content from its spatial boundary: Complementary roles for the ppa and loc in representing real-world scenes. *Journal of Neuroscience*, 2011. 1

[29] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 1, 2

[30] J. Rivest and P. Cabanagh. Localizing contours defined by more than one attribute. *Vision Research*, 1996. 2

[31] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011. 3

[32] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 4

[33] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *IJCV*, 81(1), 2007. 3, 4, 7

[34] E. Sudderth, A. Torralba, W. T. Freeman, and A. Wilsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005. 2

[35] A. Torrabla. How many pixels make an image? *Visual Neuroscience*, 2009. 1, 2

[36] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, pages 1401–1408, 2005. 2

[37] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, volume 4, pages 733–747, 2008. 2

[38] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 3, 4, 8

[39] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In *CVPR*, 2011. 1

[40] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 1, 2, 3, 4, 5, 8