Human-Machine CRFs for Identifying Bottlenecks in Scene Understanding –Supplementary Material–

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

In the supplementary material, we provide the confusion matrices for segment classification in PASCAL VOC (Figure 1). We also provide the analysis of our shape priors in the following section.

1 Analysis of Shape Prior

We first evaluate the shape priors in isolation (outside the model) to quantify how well the ground truth mask matches the shape prior mask. The pixel-wise accuracy is normalized across foreground and background. Accuracy by chance would be 50%. The accuracy is computed only using pixels that fall in the bounding box. Table 1 (top row) shows these results. The accuracy of the shape prior used in [1] ("Detector") is 72.7%. If an oracle were to pick the most accurate of the shapes (across the detector components), the accuracy would be 78.5%. We asked humans to look at the average shape masks classify them into the object categories. Human performance at this task was 60%.

We also experiment with alternative shape priors. We resize the binary object masks in the training images to 10×10 pixels. This produces a 100 dimensional vector for each mask. We use K-Means over these vectors to cluster the masks. We set the number of clusters for each category to be equal to the number of detector's components for that category to be comparable to the detector prior. The shape mask for each cluster is the binary mask that is closest to the cluster center. If an oracle were to pick the most accurate of these K masks, it would achieve an accuracy of 75.3% ("Cluster" in Table 1), not better than the detector-based oracle above. If we were to pick the shape mask from the training images (without clustering) that matches the ground truth segmentation of an object the best, we get an oracle accuracy of 88.4% ("Training Mask"). This gives us a sense of the accuracy one can hope to achieve by transferring shape masks from the training data without learning a generalization.

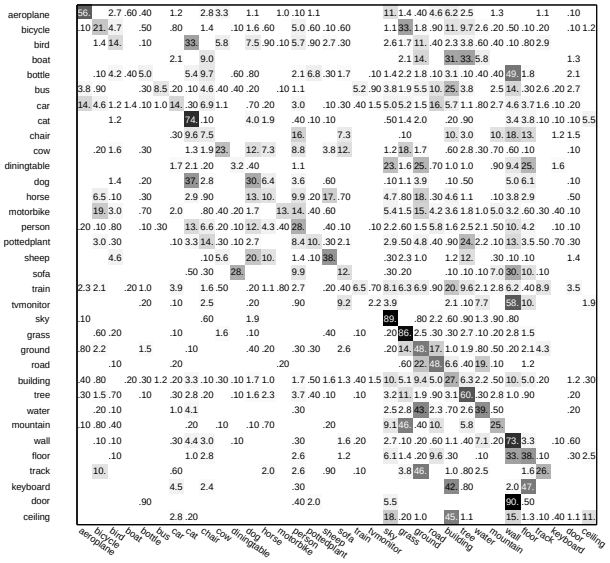
As another prior, we find the training mask whose shape matches the contours within a bounding box the best. We compute edges in the bounding boxes by thresholding the gPb contours. We compute the distance transform of these edges, and identify the training mask whose boundaries fall in regions closest to the edges. This training mask provides the shape prior for the bounding box. This *automatic* approach ("Dist. Tr.") has an accuracy of 78.8% which is comparable to the *oracle* on detector's masks.

Finally, we also experiment with a Naive approach. We simply encourage all segments that lie fully within the bounding box to take the corresponding class-label. The performance is 74.3%, higher than automatically picking the detector mask using the mixture component!. This approach shares similarities with the superpixel straddling cue of [2] which assumes that tight bounding boxes around objects do not have a lot of superpixels straddling the bounding box boundaries.

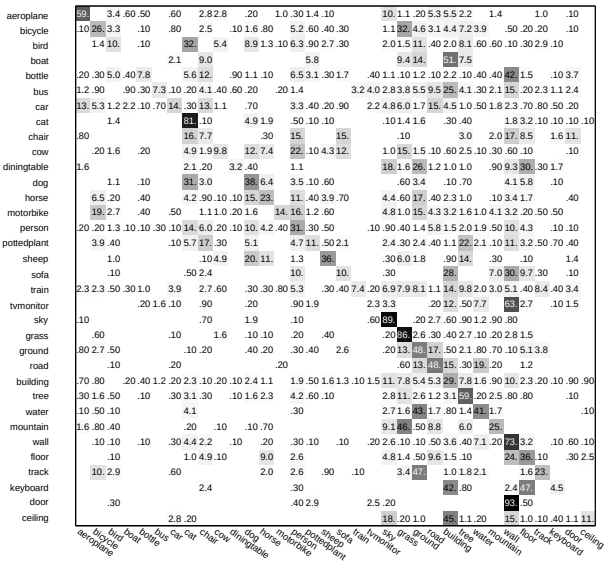
Note that if we "snap" the ground truth segmentation of an object to the segment boundaries i.e. each segment is turned on/off based on whether most of the pixels in the segment are foreground / background in the ground truth, we get an accuracy of 93.1%. This is the upper-bound on the performance given the choice of segments.

We now evaluate the impact of the various shape potentials when used in the full model. The second two rows in Table 1 provide the semantic segmentation accuracy, while bottom two rows correspond to object detection accuracy. The bottom row in both pairs of rows corresponds to using the ground truth bounding boxes around an object, while the top rows correspond to using an object detector to automatically determine the object bounding boxes.

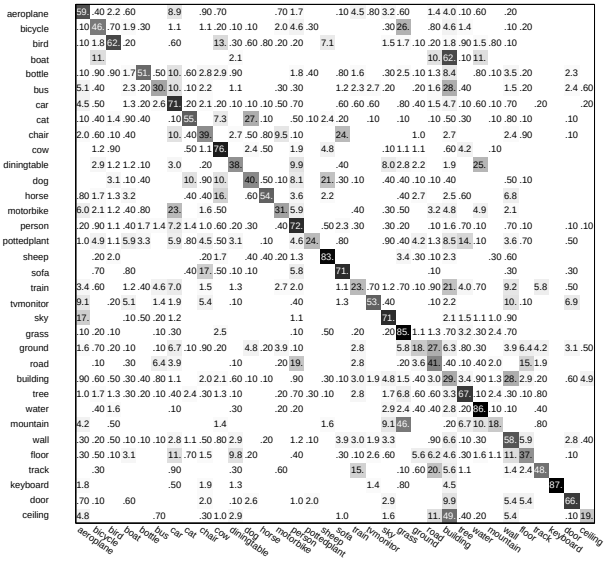
The improvement in performance we obtained over the detector induced prior used in [1] via our distance transform over contour detection approach outside the model (Table 1 top row) did not translate into an improvement in the performance of the overall model. Further analysis into this revealed that while the normalized binary segmentation accuracy for our approach was better, the unnormalized accuracy was slightly worse, which corresponds better to the metric the model is trained to optimize. While GT shape can provide significant improvement in conjunction with GT bounding boxes for objects, we find that human subjects were not able to realize this potential in terms of segmentation accuracy. Interestingly, for object detection, using human shape on ground truth detection outperforms using ground truth shape on ground truth detection.



(a) Machine classification with 100x100 patches



(b) Machine classification with 30x30 windows



(c) Human classification

Fig. 1. The confusion matrices for segment classification on PASCAL dataset. We have used 100×100 and 30×30 patches to train the machine classifier ((a) and (b) respectively). The result of human segment classification is shown in (c).

	Oracle			**Automatic**				
	Detector	Training Mask	Cluster	Detector	Dist. Tr.	Naive	**Human**	**GT**
Separate	78.5	88.4	75.3	72.7	78.8	74.3	80.2	93.1
Segmentation (Det.)	77.7	78.4	77.5	77.2	76.8	76.3	77.4	80.2
Segmentation (GT)	80.9	82.3	81.1	80.8	81.6	79.5	80.8	84.5
Object Det. (Det.)	46.8	47.6	47.7	46.8	47.4	46.7	47.2	48.5
Object Det. (GT)	95.8	90.3	95.1	93.9	93.3	94.5	96.4	92.7

TABLE 1
Accuracies of different shape priors inside and outside the model. Average recall and Average Precision is reported in the middle two rows and the bottom two rows, respectively.

REFERENCES

[1] 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] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *CVPR*, 2010. 1