Supplementary Material: 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

This document is supplementary to the paper [2] and contains the following items:

- Section 1: Human and machine co-occurrence trees
- Section 2: Per class accuracies for the plot in Figure 6
- Section 3: Human and machine confusion matrices for classification of isolated segments.

The machine model used and analyzed in the paper is [4]. Machine classification of segments in Sec.3 is obtained with Textonboost [3].

References

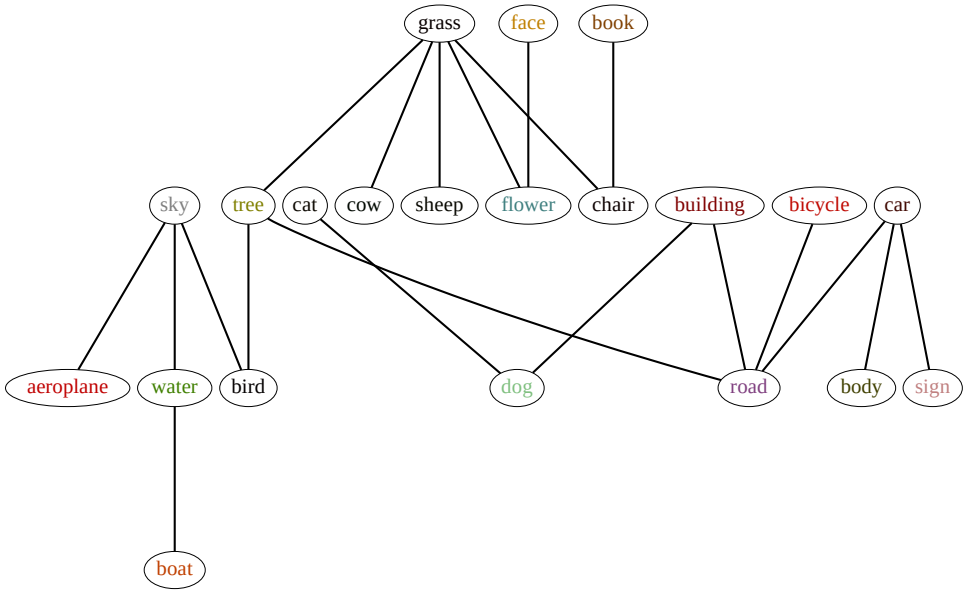
[1] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462467, 1968. 2

[2] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh. Analyzing semantic segmentation using human-machine hybrid crfs. In *CVPR*, 2013. 1

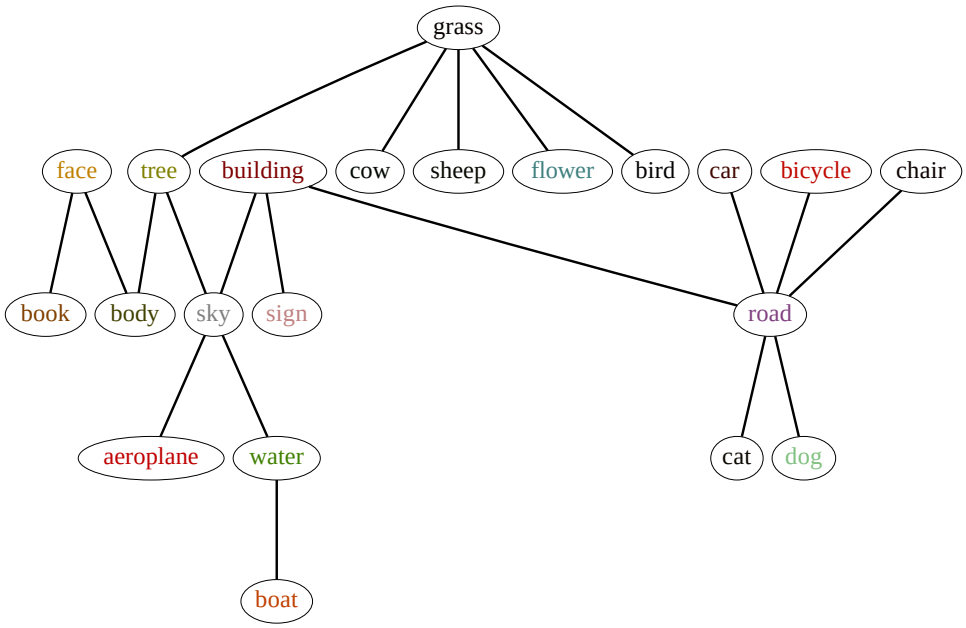
[3] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *IJCV*, 81(1), 2007. 1

[4] 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

1. Human and machine co-occurrence trees



(a) Human



(b) Machine

Figure 1. Chow-Liu trees [1] for humans and machine. The co-occurring categories are connected by edges. The machine tree is obtained using co-occurrence statistics from labeled training data from the MSRC dataset. The human tree is obtained by asking people (without showing them images) which pairs of categories are more likely to co-occur in a scene, and thus reflects the natural statistics of the world as perceived by humans. Both trees share several similarities.

2. Per class accuracies for the plot in Figure 6

In Tables 1–9, we show the per class accuracies for semantic segmentation in cases where only a single potential of the model is changed. The accuracies averaged across classes are shown in Figure 6 of the paper.

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	Segments																					
R	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
M	68	92	80	82	90	97	86	83	87	79	94	96	82	36	98	68	86	82	56	62	18	77.2
H	81	96	86	91	92	95	84	94	91	82	94	94	90	62	93	74	89	80	71	71	21	82.4
GT	**96**	**99**	**96**	**92**	**92**	**99**	**86**	**99**	**96**	**94**	**94**	**99**	**98**	**89**	**100**	**94**	**96**	**93**	**92**	**90**	**86**	**94.2**

Table 1. Segmentation accuracy for different types of segment potential.

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	Super-segments																					
R	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
M	68	92	80	82	90	97	**86**	83	87	79	94	96	82	36	98	68	86	82	56	62	18	77.2
H	**74**	**95**	83	90	90	96	84	**90**	**92**	**89**	95	97	85	**56**	99	71	**88**	81	54	71	27	81.2
GT	72	94	**86**	**91**	**91**	**98**	84	86	92	85	**95**	**97**	**93**	54	**99**	**83**	87	**85**	**59**	**72**	**32**	**82.6**

Table 2. Segmentation accuracy for different types of super-segment potential.

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	Scene																					
R	68	92	80	82	89	97	86	83	86	79	94	96	85	35	98	**70**	86	78	55	62	23	77.4
M	68	92	80	82	90	97	86	83	**87**	79	94	96	82	36	98	68	86	**82**	56	62	18	77.2
H	68	92	80	82	90	97	86	83	86	79	94	96	85	36	98	68	86	78	56	62	24	77.4
GT	**68**	**92**	**80**	**82**	**90**	**97**	**86**	**83**	86	**79**	**94**	**96**	**85**	**36**	**98**	68	**86**	79	**56**	**63**	**25**	**77.5**

Table 3. Segmentation accuracy for different types of scene potential.

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	Shape																					
R	68	92	80	83	90	97	86	83	84	81	94	97	83	30	98	68	86	82	56	62	10	76.6
M	68	92	80	82	90	97	86	83	87	79	94	96	82	36	98	68	86	82	56	62	18	77.2
H	69	92	80	85	91	97	85	83	84	80	94	96	82	30	98	67	**87**	82	55	**67**	22	77.4
GT	**69**	**92**	**80**	**91**	**91**	**97**	**86**	**83**	**93**	**84**	**94**	**97**	**83**	**36**	**98**	**73**	86	**82**	**57**	61	**49**	**80.2**

Table 4. Segmentation accuracy for different types of shape potential.

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	Class-Class																					
R	68	91	80	82	90	97	86	83	87	79	94	96	82	36	98	68	86	78	55	62	**19**	77.0
M	68	92	80	82	90	97	86	83	87	79	94	96	82	36	98	68	86	82	56	62	18	77.2
H	**68**	**92**	**80**	**82**	**90**	**97**	**86**	**83**	**87**	**79**	**94**	**96**	**82**	**36**	**98**	**68**	**86**	**82**	**56**	**62**	18	**77.2**
GT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Table 5. Segmentation accuracy for different types of class-class potential.

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	P^n **Potts**																					
R	**70**	90	**83**	**84**	**91**	**98**	85	81	**89**	77	91	93	**86**	36	98	68	**87**	**83**	56	60	10	77.0
M	68	**92**	80	82	90	97	**86**	**83**	87	**79**	**94**	**96**	82	**36**	**98**	**68**	86	82	**56**	**62**	**18**	**77.2**
H	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
GT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Table 6. Segmentation accuracy for different types of P^n Potts potential.

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	Scene-Class																					
R	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
M	68	92	80	82	**90**	97	86	83	**87**	79	94	96	82	**36**	98	68	86	**82**	**56**	62	18	77.2
H	**68**	**92**	**80**	**82**	89	**97**	**86**	**83**	86	**79**	**94**	**96**	**85**	35	**98**	**70**	**86**	78	55	**62**	**22**	**77.3**
GT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Table 7. Segmentation accuracy for different types of scene-class potential.

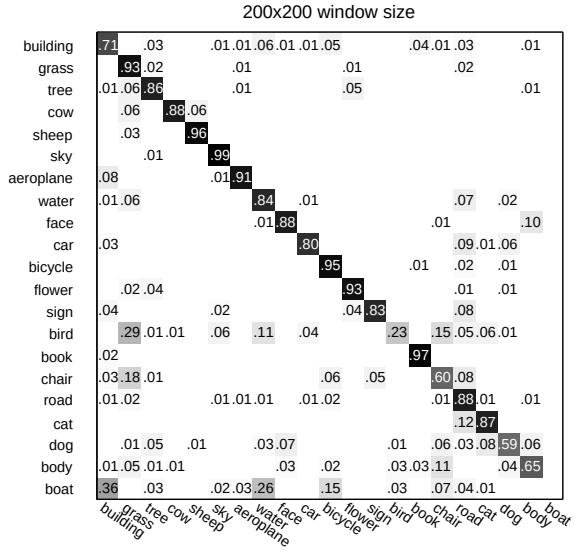
	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	Detector																					
R	67	92	**81**	81	89	**98**	84	**84**	83	78	93	96	78	30	98	66	**88**	77	54	60	0	75.0
M	**68**	92	80	82	90	97	86	83	87	79	94	96	82	**36**	98	68	86	82	56	**62**	18	77.2
H	67	92	80	83	91	97	86	83	92	84	95	98	92	32	100	71	87	83	57	61	69	80.8
GT	67	**92**	80	**83**	**91**	97	**86**	83	**92**	**84**	**95**	**98**	**92**	32	**100**	**71**	87	**83**	**57**	61	**69**	**80.8**

Table 8. Segmentation accuracy for different types of detector potential.

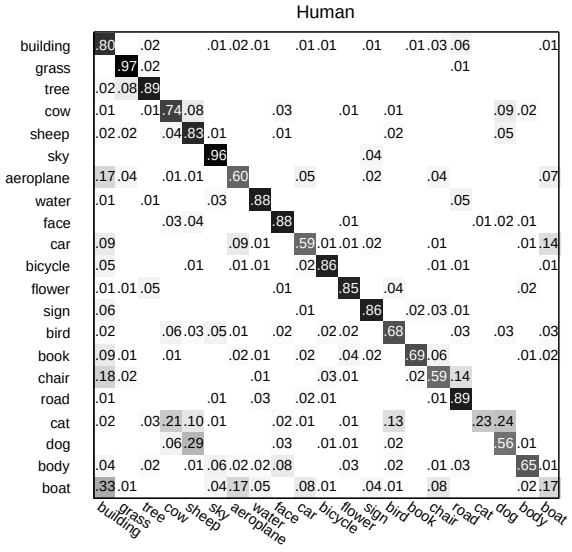
	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**
	Class Presence																					
R	68	92	80	84	90	97	86	84	87	80	**95**	96	83	35	98	72	86	82	56	62	18	77.6
M	68	92	80	82	90	97	86	83	87	79	94	96	82	36	98	68	86	82	56	62	18	77.2
H	68	91	81	85	**91**	97	86	83	**87**	80	94	96	86	35	98	**72**	84	**83**	55	61	18	77.8*
GT	**69**	**93**	**82**	**88**	90	**98**	**86**	**84**	84	**80**	94	**96**	**87**	**37**	**99**	70	**86**	79	**57**	**64**	**28**	**78.6**

Table 9. Segmentation accuracy for different types of class presence potential. *It should be noted this does NOT correspond to showing humans an image and asking them whether a certain object is present in the image or not. Human responses would be perfect and would correspond to GT. Instead, we show subjects a sampling of images from the dataset and ask them to build an intuition for the scenes. We then ask them which categories are likely to be more frequent than others in the dataset. We use these as potentials instead of the category occurrence frequencies extracted from training images.

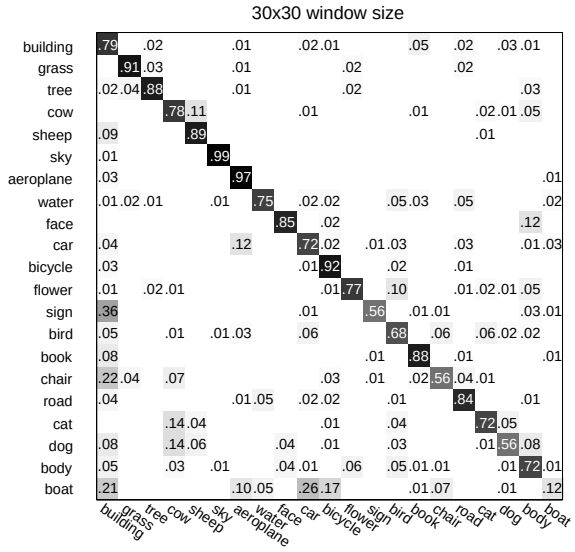
3. Human and machine confusion matrices for classification of isolated segments



(a) Machine classification with 200x200 windows



(b) Human classification



(c) Machine classification with 30x30 windows

Figure 2. 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). 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, resulting in the state-of-the-art performance on the MSRC dataset.