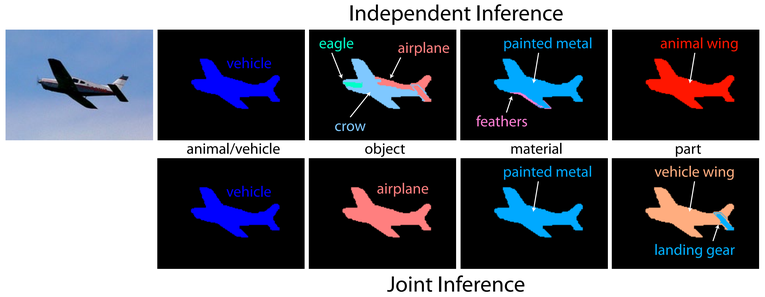
Understanding Scenes on Many Levels

Joseph Tighe and Svetlana Lazebnik Dept. of Computer Science, University of North Carolina at Chapel Hill

Overview

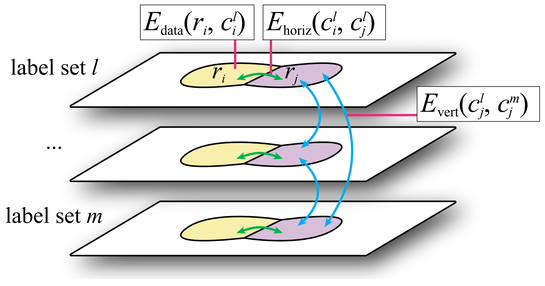
We present an image parsing system that simultaneously labels every image region according to multiple label sets, e.g., basic-level and superordinate categories, geometric classes, materials, parts, etc. By computing co-occurrence statistics of different label types for the same region, we capture relationships such as "airplanes are made of metal," "roads are horizontal," "cars have wheels" and so on. Incorporating these relationships into a joint MRF inference framework gives us a richer form of image understanding.



Multi-Level Inference

We want to infer n labelings $\mathbf{c}^1, \ldots \mathbf{c}^n$, where $\mathbf{c}^l = \{c_i^l\}$ is the vector of labels from the lth set for every region $r_i \in R$:

$$E(\mathbf{c}^1, \ldots, \mathbf{c}^n) = \sum_l \sum_{r_i \in R} E_{\text{data}}(r_i, c_i^l) + \lambda \sum_l \sum_{(r_i, r_j) \in A} E_{\text{horiz}}(c_i^l, c_j^l) + \mu \sum_{l \neq m} \sum_{r_i \in R} E_{\text{vert}}(c_i^l, c_i^m)\,.$$



E_{data} is the output of a region-based classifier ($L(r_i, c_i^l)$), passed through a sigmoid nonlinearity and scaled by the size of the region (w_i):

$$E_{\text{data}}(r_i, c_i^l) = w_i \sigma(L(r_i, c_i^l))\,.$$

E_{horiz} enforces smoothness between neighboring regions within a single label set:

$$E_{\text{horiz}} = -\log[(P(c_i|c_j) + P(c_j|c_i))/2] \times \delta[c_i \neq c_j]\,.$$

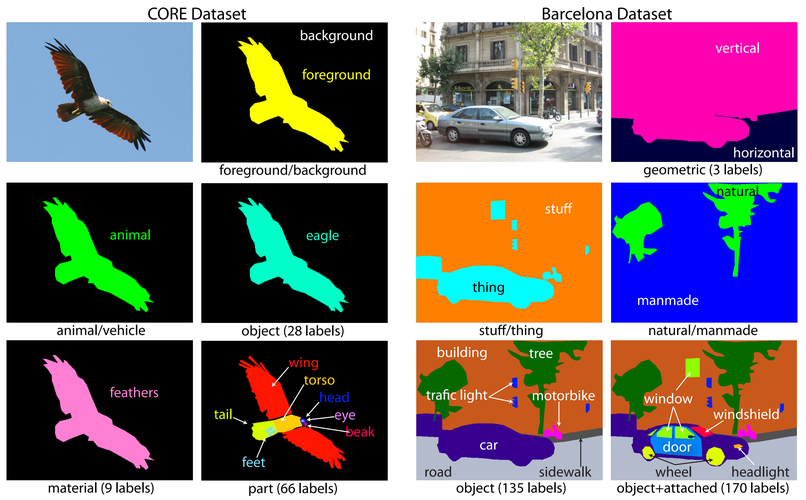
E_{vert} enforces consistency of different label types of the same region:

$$E_{\text{vert}}(c_i^l, c_i^m) = -\log[(P(c_i^l|c_i^m) + P(c_i^m|c_i^l))/2]\,.$$

Sample E_{vert} values for "object" and "part" label sets in the CORE dataset:

	bus	semi	carriage	car	motorcycle	bicycle	snowmobile	airplane	hovercraft	boat	ship	jetski	blimp	alligator	lizard	eagle	crow	bat	penguin	camel	elephant	cat	dog	monkey	elk	cow	whale	dolphin
torso	.99	1	.99	.97	.98	.99	.95	.96	.96	.96	.98	.97	.96	.16	.17	.18	.21	.20	.12	.16	.18	.13	.15	.20	.17	.13	.12	.14
leg	.98	.99	.98	.96	.97	.98	.94	.96	.95	.95	.97	.96	.95	.18	.18	.35	.23	.27	.32	.12	.13	.17	.15	.18	.16	.14	.93	.92
animal wing	.96	.96	.95	.94	.95	.95	.92	.94	.93	.94	.95	.94	.93	.94	.93	.09	.13	.07	.16	.94	.95	.94	.94	.93	.93	.93	.91	.91
ear	.90	.90	.89	.89	.89	.89	.88	.89	.88	.89	.89	.89	.88	.89	.88	.88	.87	.21	.88	.34	.14	.19	.18	.28	.29	.23	.87	.87
horn	.87	.87	.87	.86	.86	.87	.85	.86	.86	.87	.87	.86	.86	.86	.86	.86	.85	.86	.86	.86	.86	.86	.86	.86	.04	.42	.85	.85
nose	.82	.82	.82	.82	.82	.82	.81	.11	.82	.82	.82	.82	.82	.34	.45	.81	.81	.30	.81	.34	.30	.31	.25	.40	.40	.27	.45	.34
wheel	.21	.18	.07	.15	.12	.06	.31	.27	.96	.96	.98	.97	.96	.97	.96	.95	.93	.97	.95	.97	.97	.97	.97	.96	.96	.96	.93	.93
hull	.98	.98	.97	.96	.97	.97	.94	.95	.95	.06	.06	.07	.95	.96	.95	.94	.93	.96	.94	.96	.96	.96	.96	.95	.95	.95	.92	.92
windshield	.09	.21	.52	.16	.26	.94	.15	.92	.92	.92	.94	.26	.92	.93	.92	.92	.91	.93	.92	.93	.93	.93	.93	.92	.92	.92	.90	.90
cabin	.94	.09	.14	.92	.93	.93	.91	.92	.92	.25	.18	.92	.25	.92	.92	.92	.90	.92	.91	.92	.93	.93	.92	.92	.92	.92	.90	.90
windows	.07	.92	.91	.90	.91	.91	.89	.26	.23	.31	.19	.90	.90	.90	.90	.90	.89	.90	.90	.91	.91	.91	.90	.90	.90	.88	.88	.87
handlebars	.90	.91	.90	.89	.16	.13	.19	.89	.89	.89	.90	.14	.89	.89	.89	.89	.88	.89	.89	.90	.90	.90	.90	.89	.89	.89	.88	.87

Ground Truth Labels From Our Two Datasets



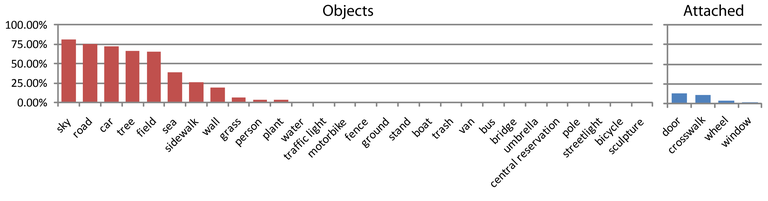
Barcelona Dataset

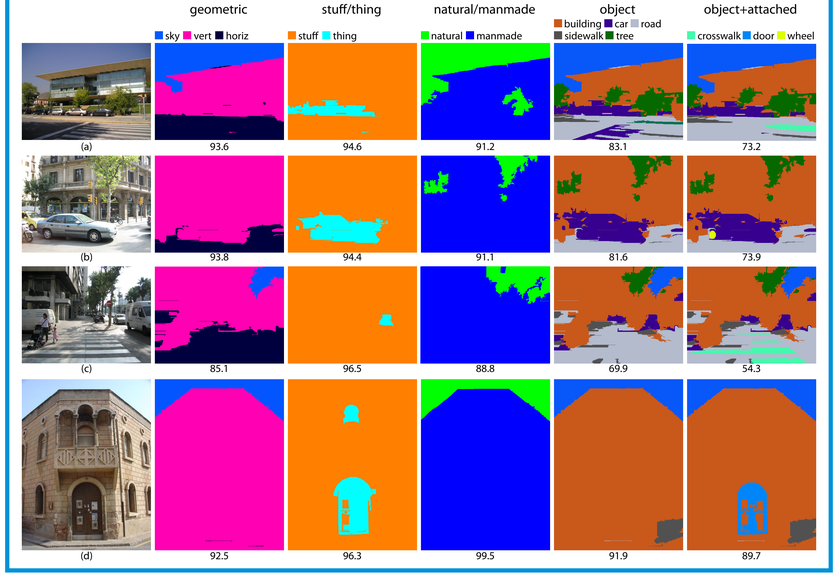
Barcelona dataset: 14,871 training images, 279 test images, 5 label sets.

Per-pixel classification rates (with average per-class rates in parentheses):

	Geometric	Stuff/thing	Natural/manmade	Objects	Objects+attached
Base (E_{data} only)	91.7 (87.6)	86.9 (66.7)	87.6 (81.8)	66.1 (9.7)	62.3(7.4)
Single-level MRF	91.4 (86.5)	89.2 (64.2)	88.4 (81.0)	68.2 (8.6)	64.4 (6.5)
Multi-level MRF	**91.8 (87.6)**	**90.3 (66.8)**	**88.9 (81.8)**	**69.3 (9.9)**	**65.2 (7.4)**
Two-level MRF	91.5 (86.8)				65.0 (7.3)

Per-class rates for stand-alone and attached objects in the Barcelona dataset:

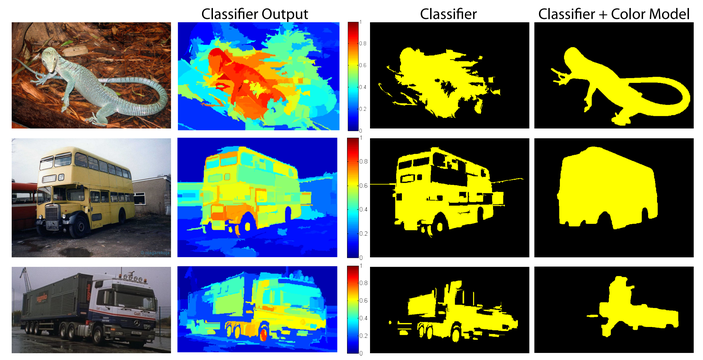




CORE Dataset

CORE (Farhadi *et al.*CVPR 2010): 2,332 training images, 224 test images, 224 validation images, 5 label sets.

CORE is an object-centric dataset, so it requires a separate foreground-background segmentation step:



We improve the segmentation using a color model and iterative refinement, similarly to GrabCut (Rother *et al.*SIGGRAPH 2004):

$$E(\mathbf{c}) = \sum_{p_i} [\alpha E_{\text{color}}(p_i, c_i) + E_{\text{data}}(p_i, c_i)] + \lambda \sum_{(p_i, p_j) \in A} E_{\text{smooth}}(c_i, c_j)\,.$$

This improves our foreground classification accuracy from 64.7% to 76.9%, and the background accuracy from 94.3% to 94.4%

Per-pixel classification rates (with average per-class rates in parentheses):

	Animal/vehicle	Object	Material	Part
Base	86.6 (86.6)	34.4 (33.4)	51.8 (36.0)	37.1 (11.2)
Single-level MRF	91.1 (91.0)	43.2 (41.7)	54.1 (34.3)	42.6 (11.7)
Multi-level MRF	91.9 (92.0)	44.5 (43.1)	54.9 (35.9)	42.7 (11.9)
Base + SVM	92.8 (92.9)	43.5 (41.8)	51.8 (36.0)	37.1 (11.2)
Single-level MRF + SVM	92.8 (92.9)	53.2 (50.5)	54.1 (34.3)	42.6 (11.7)
Multi-level MRF + SVM	92.8 (92.9)	**53.9 (51.0)**	**56.4 (36.7)**	**43.9 (12.3)**

Per-class rates for three label sets in the CORE dataset (only the top 18 parts are shown):