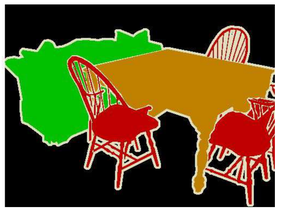
Layered Object Detection for Multi-Class Segmentation

Yi Yang, Sam Hallman, Deva Ramanan, Charless Fowlkes

Department of Computer Science, University of California, Irvine

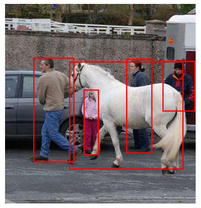
Objective

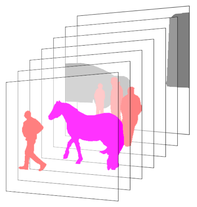


Multi-class segmentation from multi-class object detections.

- Problem 1: Need to resolve overlapping detections in order to label pixels
- Problem 2: Need to calibrate detectors relative to each other

Model





Multi-class object detection algorithms predict bounding box locations and class labels **(left)**. Multi-class segmentation algorithms provide class labels for every pixel **(center)**. We propose to use object-specific bounding box representations to guide multi-class segmentation algorithms. To do so, we introduce layered representations **(right)** that reason about relative depth orderings of objects.

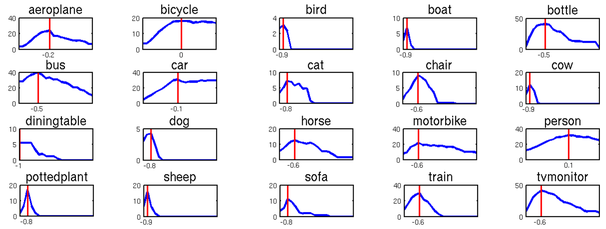
- d_n : the class, score, and bounding box coordinates of the nth detection, where $1 \leq n \leq N$.
- π : a permutation with which detections are ordered, so that $d_{\pi(N)}$ is the front-most detection.
- θ_n : the parameters of the color model associated with the nth detection.
- x_i : the color of the ith pixel.
- z_i : the label for pixel i (1...N or 0 for background)

$$P(x, z, \pi|\theta, d) = P(x, z|\theta, d_\pi)P(\pi|d)$$

$$P(x, z|\theta, d_\pi) = \prod_i P(x_i, z_i|\theta, d_\pi)$$

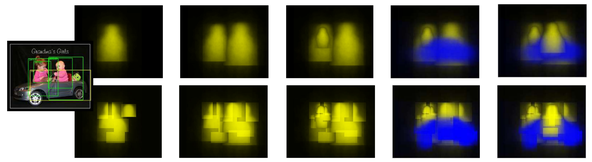
$$P(z_i = n, x_i|\theta, d_\pi) = P(z_i = n|d_\pi)\, P(x_i|\theta_n)$$

Detector calibration



- We calibrated the detector for each class independently to determine a threshold that yielded the best segmentation score. We subtracted this threshold from the detector response and kept detections which scored greater than 0.

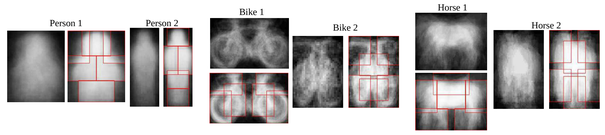
Layered label distribution



- Top row is the composited layered distribution iteratively built from detections ordered back to front.
- Bottom row is the distribution built from part-based detections which deform to better match the shape of the detected instance.

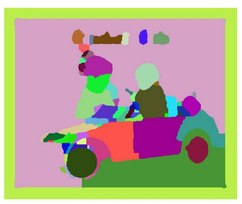
$$P(z_i = m|d_\pi) = \beta_{im} \prod_{n=m+1}^{N} (1 - \beta_{in})$$

Shape priors



- Per class shape priors are derived from a mixture model of deformable parts including both root and part templates [1].
- The mixture models capture shapes corresponding to different aspects, and part-based shape models tend to be more peaked than the root.

Bottom-up grouping



- We would like to use bottom-up grouping constraints (such as the presence of contours) to help label pixels.
- We use a segmentation engine to generate superpixels [2] and enforce the constraint that all pixels inside a superpixel must be assigned to the same layer.

Order prior

- Because local object models tend to score higher on unoccluded instances, we favor depth orderings that place high scoring objects in front.
- If multiple objects rest on a ground-plane, the object whose bottom edge is lower in the image is typically closer to the camera.

$$P(x, z, \pi|\theta, d) = P(x, z|\theta, d_\pi)P(\pi|d)$$

Putting it all together

For each ordering π, iterate until convergence:

1. $z_{S_i} := \arg\max_m \prod_{j \in S_i} P(z_j = m|d_\pi)P(x_j|\theta_m)P(\pi|d)$
2. $\theta_m(j) := \frac{\sum \mathbf{1}_{[x_i = j \text{ and } z_i = m]}}{\sum \mathbf{1}_{[z_i = m]}}$

Output superpixel labels z_{S_i} with most probable ordering π.

Results

- We analyze different components of our system on the PASCAL VOC 2009 segmentation training and validation data. The rightmost column is our full system, while the middle four represent our full system minus particular components, such as ordering, instance-specific color estimation, bottom-up grouping, and part-based priors.

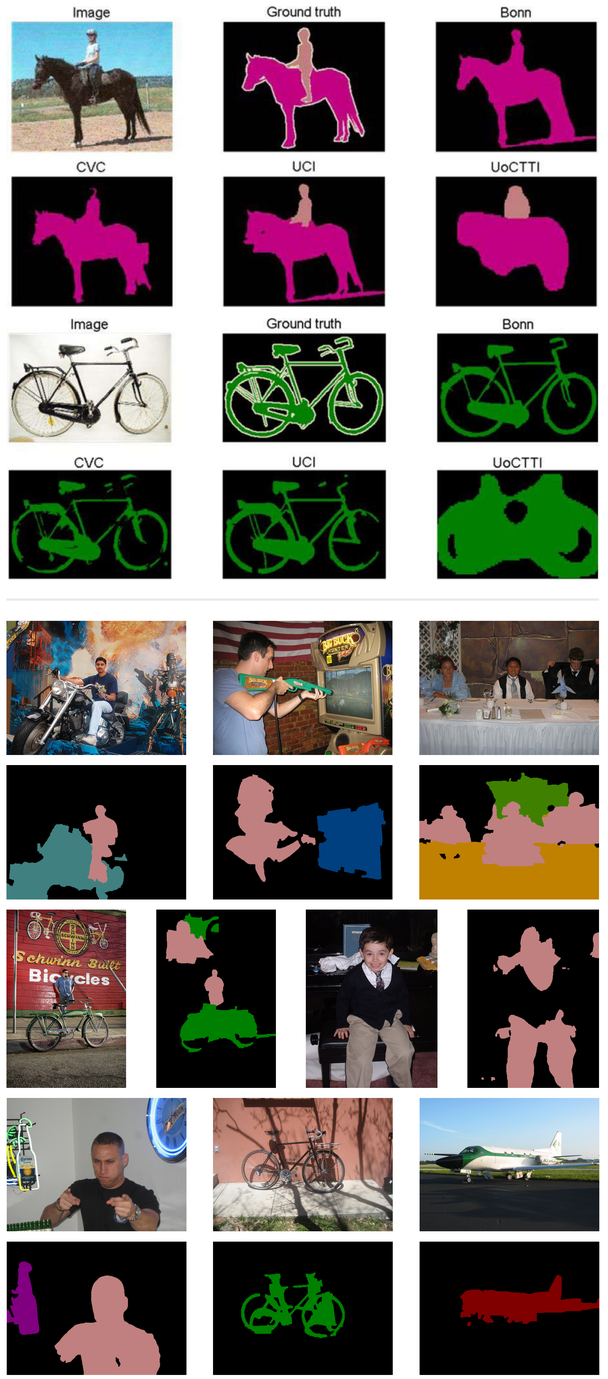
	¬ ordering	¬ color	¬ superpixel	¬ parts	all
background	79.37	78.93	78.65	79.62	79.36
aeroplane	35.26	32.39	30.61	37.22	35.26
bicycle	25.46	23.12	20.7	24.58	25.45
bird	2.81	2.78	2.68	2.79	2.81
boat	9.87	9.16	9.64	9.14	9.87
bottle	41.44	39.73	41.76	40.19	41.29
bus	49.83	48.52	48.54	48.72	49.87
car	46.88	45.66	44.25	46.14	47.03
cat	18.4	17.68	16.81	15.06	18.4
chair	10.05	9.06	9.57	8.37	10
cow	17.74	16.83	18.1	15.91	17.77
diningtable	6.94	6.8	6.79	6.85	7.27
dog	11.53	10.55	11.18	10.91	11.53
horse	16.07	14.6	15.33	15.19	16.21
motorbike	25.72	24.38	24.46	24.88	25.62
person	36.88	34.98	35.3	32.4	36.81
pottedplant	15.55	14.92	15.17	14.32	15.55
sheep	21.09	18.77	20.33	17.77	21.1
sofa	12.63	12.2	12.14	12.05	12.63
train	28.6	27.43	27.88	27.86	28.6
tvmonitor	46.41	46.01	46.36	43.67	46.28
average	26.6	25.45	25.53	25.41	26.6

- Our system performs quite well for segmenting bicycles and people. Because people represent the overwhelmingly common object in PASCAL, our system tends to produce quite reasonable segmentations overall.
- For bicycle, bottom-up grouping provides a clear improvement.
- For people, the color estimation and deformable part-based prior provides a strong improvement. This is likely because people tend to vary in appearance due to clothing, and our instance-specific color model is able to guide the final pixel labeling to more accurate configurations. Similarly people articulate their limbs, and so our part-based prior is able to better bias the grouping process.

	Mean	Max	Us	Rank
background	41.2	83.5	78.0	8
aeroplane	18.8	56.3	32.8	7
bicycle	10.4	26.6	29.4	1
bird	11.0	40.6	3.2	17
boat	11.5	36.1	5.0	16
bottle	18.2	46.1	33.1	3
bus	25.5	50.5	43.4	3
car	20.6	42.3	43.8	1
cat	12.6	35.3	8.3	12
chair	4.2	9.1	5.1	9
cow	11.7	33.1	11.9	9

	Mean	Max	Us	Rank
diningtable	9.1	27.0	8.2	11
dog	9.1	24.5	5.6	14
horse	17.5	42.7	21.0	7
motorbike	23.4	56.4	24.4	9
person	20.9	37.5	38.6	1
pottedplant	9.7	37.1	14.6	6
sheep	19.7	43.6	14.8	13
sofa	8.5	21.9	3.5	17
train	19.2	41.0	27.5	7
tvmonitor	22.3	47.8	45.7	2
average	16.4	36.2	23.7	7

Images



References

[1] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE PAMI **99** (5555)

[2] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From contours to regions: An empirical evaluation. In: CVPR. (2009)