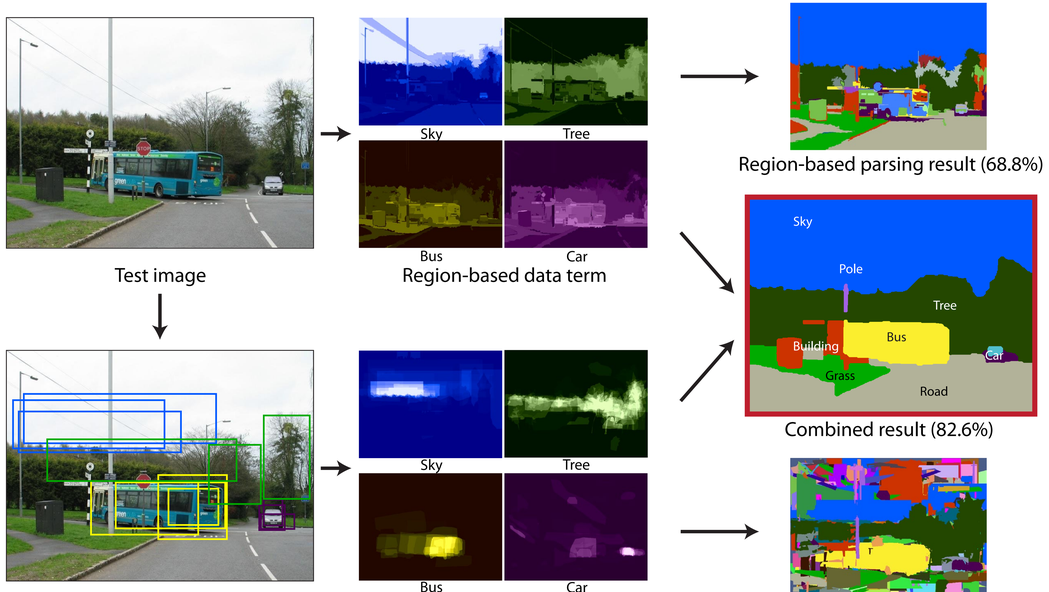
Finding Things: Image Parsing with Regions and Per-Exemplar Detectors

Joseph Tighe, University of North Carolina at Chapel Hill

Svetlana Lazebnik, University of Illinois at Urbana Champaign

Overview

We present a system for image parsing aimed at achieving broad coverage across hundreds of object categories, many of them sparsely sampled. The system combines region-level features with per-exemplar sliding window detectors. Region-level features are highly effective in identifying "stuff" categories (sky, road, building, etc.) but are quite bad at localizing "things" (car, person, sign, etc.). Conversely, sliding window detectors can reliably localize "things" but have a hard time with "stuff."



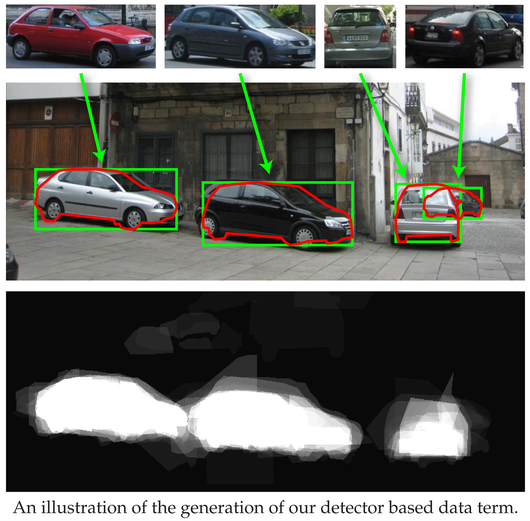
System Description

At training time, we train HOG based per-exemplar detectors (Malisiewicz et al. 2011), and compute the necessary features for our Superparsing system (Tighe and Lazebnik 2010).

Parsing pipeline:

- Obtain a retrieval set of globally similar training images
- Region based data term (E_{R}) is computed using our Superparsing system
- Detector based data term (E_{D}) :
  - Run per-exemplar detectors for exemplars in the retrieval set
  - Transfer masks from all detections above a set detection threshold to test image
  - Detector data term is computed as the sum of these masks scaled by their detection score
- Combine these two data terms by training a SVM on the concatenation of E_{D} and E_{R}
- Smooth the SVM output (E_{SVM}) using a MRF:

$$J(\mathbf{c}) = \sum_{p_i \in I} \max[0, M - E_{\mathrm{SVM}}(p_i, c_i)] + \lambda \sum_{(p_i,p_j)\in\epsilon} E_{\mathrm{smooth}}(c_i, c_j)$$



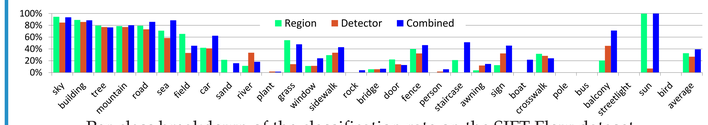
An illustration of the generation of our detector based data term.

Dataset Overview

	SIFT Flow		LM+SUN		CamVid	
	2,488 training images 200 test images 33 class labels		45,176 training images 500 test images 232 class labels		468 training images 233 test images 11 class labels	
	Per-Pixel	Per-Class	Per-Pixel	Per-Class	Per-Pixel	Per-Class
Detector ML	65.1	25.8	33.0	14.1	61.2	45.5
Detector SVM	62.5	25.4	46.1	12.0	61.4	47.0
Detector SVM MRF	71.1	26.7	52.5	11.3	63.8	47.3
Region ML	74.1	30.2	51.5	7.5	82.7	51.2
Region SVM	75.0	35.9	56.3	6.7	81.4	55.7
Region SVM MRF	77.7	32.8	58.3	5.9	83.5	55.7
Reg.+Thing SVM	74.4	36.9	58.5	14.1	82.4	60.0
Reg.+Thing SVM MRF	77.5	35.7	60.0	12.9	84.2	59.5
Combined SVM	75.6	**41.1**	59.6	**15.5**	82.3	62.1
Combined SVM MRF	**78.6**	39.2	**61.4**	15.2	**84.0**	**62.2**

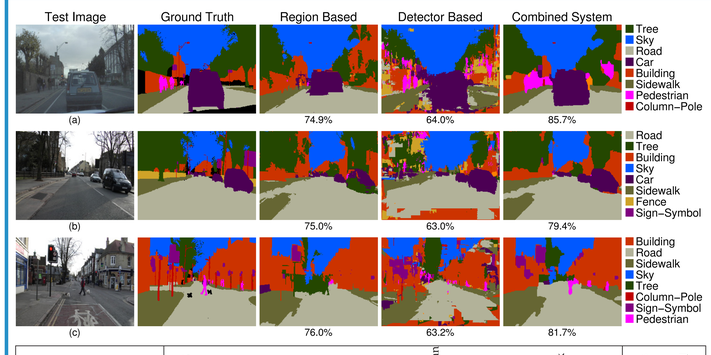
SIFT Flow Dataset

SIFT Flow	Per-Pixel	Per-Class
Combined MRF	78.6	39.2
Tighe and Lazebnik 2013	77.0	30.1
Liu et al. 2011	76.7	N/A
Farabet et al. 2012	78.5	29.6
Farabet et al. 2012 balanced	74.2	46.0
Eigen and Fergus 2012	77.1	32.5
Myeong et al. 2012	77.1	32.3



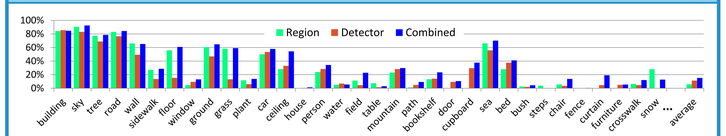
Per-class breakdown of the classification rate on the SIFT Flow dataset.

CamVid Dataset

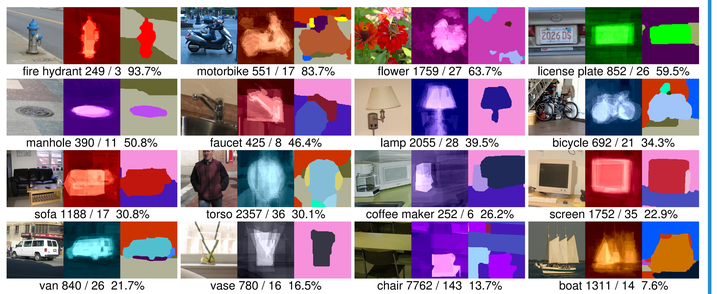


	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Bicyclist	Per-class	Per-pixel
Combined MRF	83.1	73.5	94.6	78.1	**48**	96	**58.6**	32.8	5.3	71.2	**45.9**	**62.5**	**83.9**
Tighe and Lazebnik 2013	**87.0**	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70.0	19.4	51.2	83.3
Brostow et al. 2008	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1
Sturgess et al. 2009	84.5	72.6	**97.5**	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8
Zhang et al. 2010	85.3	57.3	95.4	69.2	46.5	**98.5**	23.8	44.3	**22.0**	38.1	28.7	55.4	82.1
Floros et al. 2012	80.4	76.1	96.1	86.7	20.4	95.1	47.1	47.3	8.3	79.1	19.5	59.6	83.2
Ladicky et al. 2010	81.5	**76.6**	96.2	**78.7**	40.2	93.9	43.0	**47.6**	14.3	**81.5**	33.9	**62.5**	83.8

LM+Sun Dataset



Breakdown of the classification rate on the most common classes in the LM+SUN dataset.



Examples of "thing" classes on LM+SUN. The caption for each class shows: (# of training instances of that class) / (# of test instances) (per-pixel rate on the test set)%