Scene-Decomposition

April 12, 2013

Abbreviation:	scene-decomposition
Number of instances:	715
Number of variables:	100-200
Number of labels:	8
Number of factors:	500-1000
Order:	2
Function type:	explicit

Description The task is to segment a natural image into eight semantic categories. The original source of the data set is [1], and the data set was obtained from the Stanford DAGS group website.[1]

The standard evaluation procedure on this data set is five-fold cross validation and I trained five factor graph models on superpixel graphs. The instances included are the respective test sets, totalling 715 instances. The superpixelizations have been created using superpixels1.1 [3].

Because the original data set includes a few unlabeled pixels, a simple preprocessing step is performed to assign each unlabeled pixel the nearest spatial label in the ground truth. In addition, because we use superpixels, we assign to the superpixel the ground truth label that appears most frequently in all the pixels covered by the superpixel segment.

Objective / Learning Each superpixel factor graph is composed of a set of superpixels V and has unary and pairwise factors composing an overall energy as follows.

$$J(x; w, I) = \sum_{i \in V} \langle w_u(x_i), F_u(I, i) \rangle + \sum_{(i,j) \in \mathcal{E}} \langle w_p(x_i, x_j), F_p(I, i, j) \rangle,$$

where $\mathcal{E} \subset V \times V$ is the set of adjacent superpixels and w and F are parameters and feature maps, respectively. All terms in the energy depend on the observed image and therefore the above energy is a conditional random field.

The unary feature map F_u is composed of the following image features per superpixel: SIFT bag-of-words histograms (512 dimensions), gradient histogram (24 dimensions), k-means quantized color histograms (195 dimensions), and histograms of spatial location (42 dimensions). Together these are 773 feature dimensions for each superpixel and there is one weight vector w_u for each class label $x_i \in \{1, 2, \dots, 8\}$, hence there are $773 \cdot 8 = 6184$ parameters in this factor type.

Each pair of superpixels that is adjacent in the image plane has a data-dependent pairwise factor, again with linear energy parametrization. The features $F_p(I, i, j)$ encode a number of image-dependent pairwise properties in a low-dimensional vector: mean gradient magnitude along the edge of the two superpixels, symmetrized KL-divergence between two multivariate Normals describing the color distribution in both superpixels, and a quantized angle between the line connecting the centers of the two superpixel and the horizon. The feature vector has a dimension of 40, for a total of 8*8*40=2560 pairwise parameters.

The total number of model parameters is 8744. For each of the five models from each cross validation fold we estimate the parameters using regularized maximum pseudolikelihood, using L-BFGS numerical minimization for 750 iterations, taking less than one hour each for training. Because the energy is linear in the parameters this is a convex optimization problem. We place priors on the unary factors (multivariate Normal, $\sigma = 100$), and the pairwise factors (multivariate Normal, $\sigma = 0.01$). The MAP predictions of these models achieve 77.41% accuracy (MPM predictions 77.35%), whereas the best reported literature result is 79.42% in [2]. A model without pairwise factors but identical unary factors achieves 75.2% accuracy and this indicates that almost all the predictive performance comes from having good

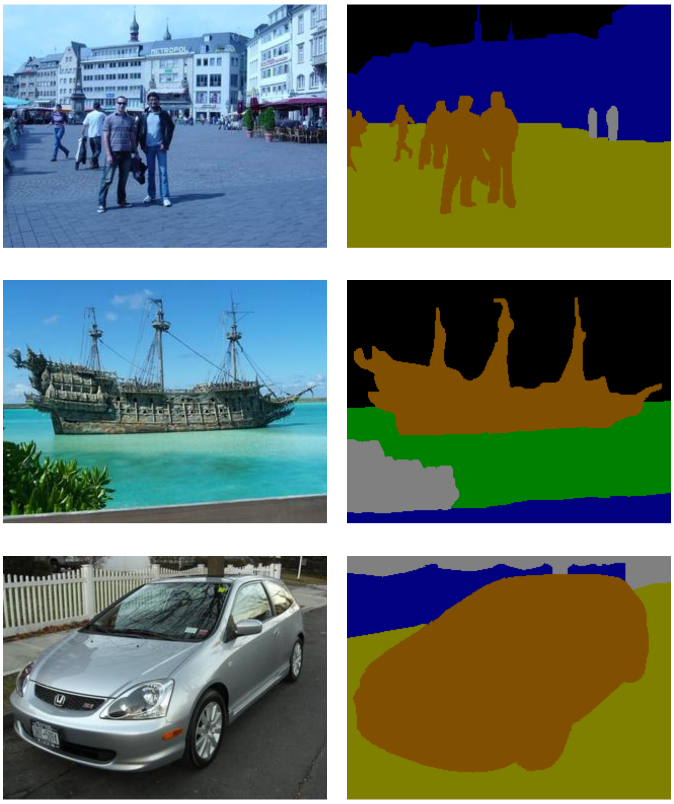

Figure 1: Three images with ground truth annotation. The orange label is the foreground object class.

[1] http://dags.stanford.edu/projects/scenedataset.html

superpixel image features.

The energy minimization instances are obtained by evaluating the effective energies for each factor instance, thus the original features and learned weights are not included with the instances.

References

[1] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.

[2] M. Pawan Kumar and Daphne Koller. Efficiently selecting regions for scene understanding. In *CVPR*, pages 3217–3224, 2010.

[3] Olga Veksler, Yuri Boykov, and Paria Mehrani. Superpixels and supervoxels in an energy optimization framework. In *ECCV (5)*, pages 211–224, 2010.