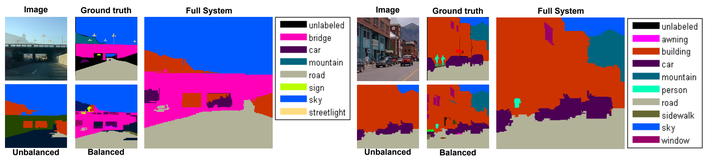
Image Parsing with a Wide Range of Classes and Scene-Level Context

Marian George

ETH zürich

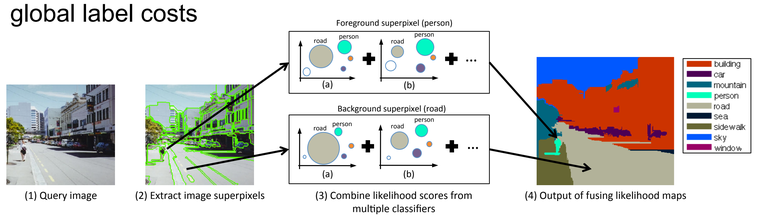
Overview

Scene parsing is the assignment of semantic labels to each pixel in a scene image



We present a nonparametric scene parsing approach that improves the overall accuracy, as well as the coverage of foreground classes in scene images:

- improve the label likelihood estimates at superpixels by merging likelihoods from different classifiers
- incorporate semantic context in the parsing process through global label costs



Baseline Parsing Pipeline

The baseline parsing system is based on (Tighe and Lazebnik 2010) but without using an image retrieval set

a) Segmentation and Feature Extraction

- extract superpixels from images
- compute 21 types of local features

b) Label Likelihood Estimation

compute a log-likelihood score for each class label c in all classes C in the dataset (no filtering):

$$L_{unbal}(s_i, c) = \tfrac{1}{2} log(P(s_i|c)/P(s_i|\bar{c}))$$

$L_{unbal}(s_i, c)$ is computed from counts in the training data

c) Smoothing and Inference

estimate the initial labeling through Markov Random Field (MRF) inference:

$$E(L) = \sum_{s_i \in S} D(l_{s_i} = c|s_i) + \lambda \sum_{(i,j)\in A} V(l_{s_i}, l_{s_j})$$

minimizing the data cost $D(l_{si} = c \mid s_i)$ and the smoothing cost $V(l_{si}, l_{sj})$

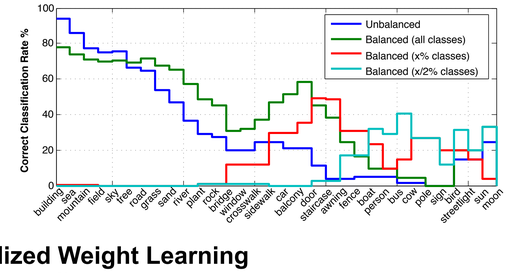
Improving Superpixel Label Costs

a) Fusing Classifiers

- combine likelihood scores from multiple classifiers to improve the overall classification accuracy
- fusing classifiers performs well when the error of individual classifiers is uncorrelated
- classification error related to mean number of pixels occupied by a class in scene images (x%)
- combine 4 classification models

$$L_{comb}(s_i, c) = \sum_{j=1,2,3,4} w_j(c)\, L_j(s_i, c)$$

$\omega_j(c)$ is the normalized weight of the score of c in the j^{th} classifier



b) Normalized Weight Learning

learn weights offline as normalized likelihood ratio:

$$\tilde{w}_j(c) = \frac{|C_j|}{C} \frac{\sum_{s_i \in S} L_j(s_i, c)}{\sum_{s_i \in S} \sum_{c_i \in C \setminus c} L_j(s_i, c_i)}$$

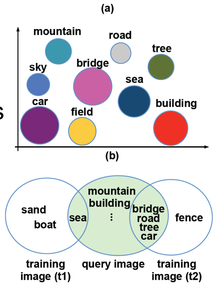
Scene-Level Global Context

We do not limit the number of labels to those present in the retrieval set.

Context-Aware Global Label Costs

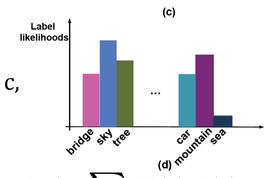
a) given the the initial labeling of an image L,
b) compute weights for unique labels T in L
c) rank images by weighted intersection of class labels with query image
d) compute global likelihood of labels in k-NN fashion:

$$P(c|T) = \frac{(1 + n(c, K_T))/n(c, S)}{(1 + n(\bar{c}, K_T))/|S|}$$



Inference with Label Costs

- define H(c) as the global label cost of label c and δ(c) as the indicator function of c,
- our final energy function becomes:



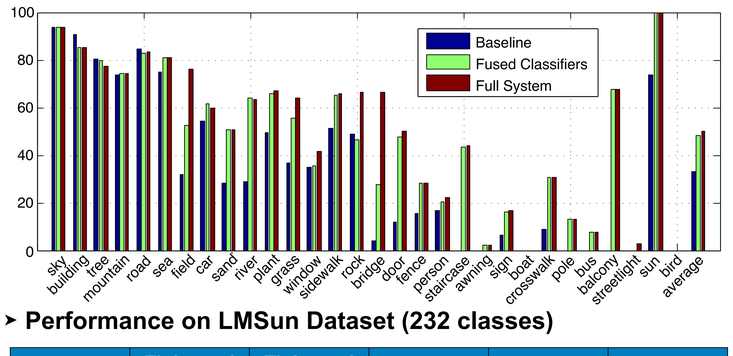
$$E(L) = \sum_{s_i \in S} D(l_{s_i} = c|s_i) + \lambda \sum_{(i,j)\in A} V(l_{s_i}, l_{s_j}) + \sum_{c \in C} H(c).\delta(c)$$

Results

Performance on SIFTflow Dataset (33 classes)

Method	Liu et al.	Farabet et al.	Farabet et al. balanced	Eigen et al.	Singh et al.	Tighe and Lazebnik, 2010	Tighe and Lazebnik, 2013	Yang et al.	Ours (FC only)	Ours (Full)
Per-pixel Accuracy(%)	76.7	78.5	74.2	77.1	79.2	77.0	78.6	79.8	80.5	81.7
Per-class Accuracy(%)	N/A	29.5	46.0	32.5	33.8	30.1	39.2	48.7	48.2	50.1

Classification rates of individual classes on SIFTflow



Performance on LMSun Dataset (232 classes)

Method	Tighe and Lazebnik, 2010	Tighe and Lazebnik, 2013	Yang et al.	Ours (FC only)	Ours (Full)
Per-pixel Accuracy(%)	54.9	61.4	60.6	60.0	61.2
Per-class Accuracy(%)	7.1	15.2	18.0	14.2	16.0