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

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):

\[ L_{balo}(s_i, c) = \frac{1}{2} \log(P(s_i|c)/P(s_i)) \]

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

c) Smoothing and Inference
estimate the initial labeling through Markov Random Field (MRF) inference:

\[ E(L) = \sum_{s_i \in S} D(l_{si} = c|s_i) + \lambda \sum_{(i,j) \in A} V(l_{si}, l_{sj}) \]

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

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 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:

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

Inference with Label Costs
- define \( H(c) \) as the global label cost of label \( c \) and \( \delta(c) \) as the indicator function of \( c \)
- our final energy function becomes:

\[ E(L) = \sum_{s_i \in S} D(l_{si} = c|s_i) + \lambda \sum_{(i,j) \in A} V(l_{si}, l_{sj}) + \sum_{c \in C} H(c) \delta(c) \]

Results

Performance on SIFTFlow Dataset (33 classes)

<table>
<thead>
<tr>
<th>Method</th>
<th>Tighe and Lazebnik, 2010</th>
<th>Tighe and Lazebnik, 2013</th>
<th>Yang et al.</th>
<th>Ours (FC only)</th>
<th>Ours (Full)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per-class Accuracy(%)</td>
<td>54.9</td>
<td>61.4</td>
<td>60.6</td>
<td>60.0</td>
<td>61.2</td>
</tr>
<tr>
<td>Per-class Accuracy(%)</td>
<td>7.1</td>
<td>15.2</td>
<td>18.0</td>
<td>14.2</td>
<td>16.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Tighe and Lazebnik, 2010</th>
<th>Tighe and Lazebnik, 2013</th>
<th>Yang et al.</th>
<th>Ours (FC only)</th>
<th>Ours (Full)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per-class Accuracy(%)</td>
<td>76.1</td>
<td>79.5</td>
<td>74.2</td>
<td>77.4</td>
<td>78.6</td>
</tr>
<tr>
<td>Per-class Accuracy(%)</td>
<td>3.3</td>
<td>3.8</td>
<td>3.5</td>
<td>3.6</td>
<td>3.7</td>
</tr>
<tr>
<td>Per-class Accuracy(%)</td>
<td>29.5</td>
<td>46.0</td>
<td>32.5</td>
<td>35.1</td>
<td>48.7</td>
</tr>
<tr>
<td>Per-class Accuracy(%)</td>
<td>20.5</td>
<td>61.7</td>
<td>21.2</td>
<td>24.7</td>
<td>50.1</td>
</tr>
</tbody>
</table>