Context Driven Scene Parsing with Attention to Rare Classes

Jimei Yang
UC Merced

Brian Price
Adobe Research

Scott Cohen
Adobe Research

Ming-Hsuan Yang
UC Merced
Scene Parsing

query

labeled database

result
An Exemplar Based Approach

- Scale up to large sets of semantic labels
- Preserve scene context
- Adapt to the database changes

Related work
- J. Tighe and S. Lazebnik, ECCV10, CVPR13
- C. Liu, J. Yuen and A. Torralba, PAMI11
- D. Eigen and R. Fergus, CVPR12
- G. Singh and J. Kosecka, CVPR13
Problem

Image

Result

- field
- mountain
- rock
- sand
- sea
- sky
- tree
Problem

Missing small objects

Image

Result

- Per-pixel accuracy = the percent of correctly labeled pixels in total
- Per-class accuracy = the average percent of correctly labeled pixels in each class
How can the system recognize these small objects?

- Per-pixel accuracy = the percent of correctly labeled pixels in total
- Per-class accuracy = the average percent of correctly labeled pixels in each class
Baseline Algorithm

- Image Retrieval
- Superpixel Matching
- MRF Labeling
- The same baseline as these papers, with slight variation
  - J. Tighe and S. Lazebnik, ECCV10, CVPR13
  - D. Eigen and R. Fergus, CVPR12
  - G. Singh and J. Kosecka, CVPR13
Baseline Algorithm: Image Retrieval

A bag-of-words image retrieval system*

* X. Shen, Z. Lin, J. Brandt, S. Avidan and Y. Wu, CVPR12
Baseline Algorithm: Image Retrieval
Baseline Algorithm: Superpixel Matching

* P. Felzenszwalb and D. Huttenlocher, IJCV04

Superpixels

database

Retrieval set with labeled superpixels
Baseline Algorithm: Superpixel Matching

Superpixels

Likelihood maps

database

Retrieval set with labeled superpixels
Baseline Algorithm: MRF Labeling

Superpixels → Likelihood maps (unary term) → Pairwise term → MRF labeling

Retrieval set

database
Observations in Image Retrieval

Label distribution in the retrieval set

Retrieval set with labeled superpixels
Insufficient Samples

Label distribution in the retrieval set

Insufficient samples: boat and person
Out-Of-Context Noise

Label distribution in the retrieval set

Outlier: Building
Problems

- Retrieval set with labeled superpixels
  - Insufficient samples: boat and person
  - Data noise

Label distribution in the retrieval set
Our Approach

• Rare class expansion
• Context feedback
Our Approach

- Rare class expansion
- Context feedback
Rare Classes
Defining Rare Classes by Pareto Principle

5 common classes
building, tree, mountain, sky, sea

28 rare classes
road, field, plant, car, grass, river, window, rock, sand, sidewalk, desert, door, bridge, fence, person, balcony, staircase, sign, awning, crosswalk, boat, streetlight, bus, pole, sun, cow, moon, bird

80% SIFT flow database

20%
Identifying Rare Classes in the Retrieval Set

sea, sky, mountain, tree, building, road, field, plant, grass, river, rock, sand, person, boat

Retrieval Set
Identifying Rare Classes in the Retrieval Set

Retrieval Set

- sea, sky, mountain, tree, building
- road, field, plant, grass, river, rock, sand, person, boat
Expanding the Rare Classes

Retrieval Set

sea, sky, mountain, tree, building

road, field, plant, grass, river, rock, sand, person, boat

Rare class expansion

boat sand field

person road river

rock plant grass
Mining Superpixel Exemplars of Rare Classes

All the boat images

boat

K-means

1000 exemplars

boat
Keyword Based Superpixel Retrieval

Road  Field  Plant  Grass  River  Rock  Sand  Person  Boat

database

boat  sand  field
person  road  river
rock  plant  grass
Rare Class Expansion

Label distribution before expansion
Rare Class Expansion

Label distribution after expansion
Computing Likelihood Maps

- Color, SIFT and HOG based superpixel descriptors
- KNN + SVM classifier
- Convert superpixel likelihoods as pixel likelihood maps
Comparisons

- Rare class expansion
- Feedback based context filtering

<table>
<thead>
<tr>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Tighe and S. Lazebnik, CVPR13</td>
</tr>
<tr>
<td>Shape information within retrieval set</td>
</tr>
<tr>
<td>Our approach</td>
</tr>
<tr>
<td>Additional information out of retrieval set</td>
</tr>
</tbody>
</table>
Irrelevant rare classes, grass, road and river become a new source of noise.
Context Filtering by Feedback

Filtering by Feedback

Superpixels

Likelihood maps

Context Complementary to appearance features

database
Context Filtering by Feedback

Context feedback

Superpixels

Likelihood maps

database

Context Complementary to appearance features
Global Context for Image Retrieval

- Combine Bag-of-Words and global context descriptors for image retrieval*

* G. Singh and J. Kosecka, CVPR13
Local Context for Superpixel Matching

- Combine appearance and local context descriptors for superpixel classification*

* Z. Tu and X. Bai, PAMI10
Comparisons

- Rare class expansion
- Context feedback

<table>
<thead>
<tr>
<th></th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Eigen and R. Fergus,</td>
<td>Local hard label distribution</td>
</tr>
<tr>
<td>CVPR12</td>
<td></td>
</tr>
<tr>
<td>G. Singh and J. Kosecka,</td>
<td>Global hard label distribution</td>
</tr>
<tr>
<td>CVPR13</td>
<td></td>
</tr>
<tr>
<td>Our approach</td>
<td>Local/global soft likelihood</td>
</tr>
<tr>
<td></td>
<td>distribution</td>
</tr>
</tbody>
</table>

- Soft likelihood maps are more informative than hard labels
- Do not require MRF inference
Full Algorithm
Full Algorithm

1. Database
2. Superpixels
3. Likelihood maps
4. Rare class expansion
5. Retrieval
Full Algorithm

Context feedback

Superpixels

New likelihood maps

database

Retrieval

Rare class expansion

Full Algorithm
Full Algorithm

Context feedback

Superpixels

New likelihood maps

Pairwise term

MRF labeling

database

Retrieval

Rare class expansion

Full Algorithm

Pairwise term

MRF labeling

Superpixels

New likelihood maps

Context feedback

database

Retrieval

Rare class expansion

Full Algorithm

Pairwise term

MRF labeling

Superpixels

New likelihood maps

Context feedback

database

Retrieval

Rare class expansion
Results on SIFT Flow

- 2488 training images and 200 test images and the images are 256 x 256 pixels from 33 semantic labels.
- 28 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>78.0</td>
<td>27.5</td>
</tr>
<tr>
<td>Baseline + rare class expansion</td>
<td>78.4</td>
<td>45.4</td>
</tr>
<tr>
<td>Baseline + rare class expansion + context feedback (our full system)</td>
<td>79.8</td>
<td>48.7</td>
</tr>
</tbody>
</table>

- Per-pixel accuracy = the percent of correctly labeled pixels in total
- Per-class accuracy = the average percent of correctly labeled pixels in each class
Results on SIFT Flow

- 2488 training images and 200 test images and the images are 256 x 256 pixels from 33 semantic labels.

- 28 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>78.0</td>
<td>27.5</td>
</tr>
<tr>
<td>Baseline + rare class expansion</td>
<td>78.4</td>
<td>45.4</td>
</tr>
<tr>
<td>Baseline + rare class expansion + context feedback (our full system)</td>
<td><strong>79.8</strong></td>
<td><strong>48.7</strong></td>
</tr>
</tbody>
</table>

- Per-pixel accuracy = the percent of correctly labeled pixels in total
- Per-class accuracy = the average percent of correctly labeled pixels in each class
Results on SIFT Flow

- 2488 training images and 200 test images and the images are 256 x 256 pixels from 33 semantic labels.

- 28 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al PAMI11</td>
<td>76.7</td>
<td>N/A</td>
</tr>
<tr>
<td>Tighe et al IJCV12</td>
<td>77.0</td>
<td>30.1</td>
</tr>
<tr>
<td>Eigen et al CVPR12</td>
<td>77.1</td>
<td>32.5</td>
</tr>
<tr>
<td>Singh et al CVPR13</td>
<td>79.2</td>
<td>33.8</td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>78.6</td>
<td>39.2</td>
</tr>
<tr>
<td><strong>Our full system</strong></td>
<td><strong>79.8</strong></td>
<td><strong>47.3</strong></td>
</tr>
<tr>
<td><strong>28 Rare Classes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>48.8</td>
<td>29.9</td>
</tr>
<tr>
<td><strong>Our full system</strong></td>
<td><strong>59.4</strong></td>
<td><strong>41.9</strong></td>
</tr>
</tbody>
</table>
Results on SIFT Flow

- 2488 training images and 200 test images and the images are 256 x 256 pixels from 33 semantic labels.
- 28 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al PAMI11</td>
<td>76.7</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Tighe et al IJCV12</td>
<td>77.0</td>
<td></td>
<td>30.1</td>
</tr>
<tr>
<td>Eigen et al CVPR12</td>
<td>77.1</td>
<td></td>
<td>32.5</td>
</tr>
<tr>
<td>Singh et al CVPR13</td>
<td>79.2</td>
<td></td>
<td>33.8</td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>78.6</td>
<td></td>
<td>39.2</td>
</tr>
<tr>
<td><strong>Our full system</strong></td>
<td><strong>79.8</strong></td>
<td></td>
<td><strong>47.3</strong></td>
</tr>
<tr>
<td>28 Rare Classes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>48.8</td>
<td></td>
<td>29.9</td>
</tr>
<tr>
<td><strong>Our full system</strong></td>
<td><strong>59.4</strong></td>
<td></td>
<td><strong>41.9</strong></td>
</tr>
</tbody>
</table>
Results on SIFT Flow

- 2488 training images and 200 test images and the images are 256 x 256 pixels from 33 semantic labels.
- 28 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al PAMI11</td>
<td>76.7</td>
<td>N/A</td>
</tr>
<tr>
<td>Tighe et al IJCV12</td>
<td>77.0</td>
<td>30.1</td>
</tr>
<tr>
<td>Eigen et al CVPR12</td>
<td>77.1</td>
<td>32.5</td>
</tr>
<tr>
<td>Singh et al CVPR13</td>
<td>79.2</td>
<td>33.8</td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>78.6</td>
<td>39.2</td>
</tr>
<tr>
<td>Our full system</td>
<td>79.8</td>
<td>47.3</td>
</tr>
<tr>
<td>28 Rare Classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>48.8</td>
<td>29.9</td>
</tr>
<tr>
<td>Our full system</td>
<td>59.4</td>
<td>41.9</td>
</tr>
</tbody>
</table>
Results on SIFT Flow

Image

Human annotation

Tighe, CVPR13

Ours
Results on SIFT Flow
Results on SIFT Flow

Image

Human annotation

Tighe, CVPR13

Ours
Results on LabelMe + SUN (LMSun)

- 45176 training images and 500 test images and the size of images ranges from 256x256 pixels to 800x600 pixels in 232 semantic labels.

- 185 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>58.5</td>
<td>9.0</td>
</tr>
<tr>
<td>Baseline + rare class expansion</td>
<td>57.1</td>
<td>14.5</td>
</tr>
<tr>
<td>Baseline + rare class expansion + context feedback (our full system)</td>
<td>60.6</td>
<td>18.0</td>
</tr>
</tbody>
</table>

- Per-pixel accuracy = the percent of correctly labeled pixels in total
- Per-class accuracy = the average percent of correctly labeled pixels in each class
Results on LabelMe + SUN (LMSun)

- 45176 training images and 500 test images and the size of images ranges from 256x256 pixels to 800x600 pixels in 232 semantic labels.

- 185 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>58.5</td>
<td>9.0</td>
</tr>
<tr>
<td>Baseline + rare class expansion</td>
<td>57.1</td>
<td>14.5</td>
</tr>
<tr>
<td>Baseline + rare class expansion + context feedback</td>
<td>60.6</td>
<td>18.0</td>
</tr>
<tr>
<td>(our full system)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Per-pixel accuracy = the percent of correctly labeled pixels in total
- Per-class accuracy = the average percent of correctly labeled pixels in each class
Results on LabelMe + SUN (LMSun)

- 45176 training images and 500 test images and the size of images ranges from 256x256 pixels to 800x600 pixels in 232 semantic labels.

- 185 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>58.5</td>
<td>9.0</td>
</tr>
<tr>
<td>Baseline + rare class expansion</td>
<td>57.1</td>
<td>14.5</td>
</tr>
<tr>
<td>Baseline + rare class expansion + context feedback (our full system)</td>
<td>60.6</td>
<td>18.0</td>
</tr>
</tbody>
</table>

- Per-pixel accuracy = the percent of correctly labeled pixels in total
- Per-class accuracy = the average percent of correctly labeled pixels in each class
Results on LabelMe + SUN (LMSun)

- 45176 training images and 500 test images and the size of images ranges from 256x256 pixels to 800x600 pixels in 232 semantic labels.

- 185 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tighe et al IJCV12</td>
<td></td>
<td>54.9</td>
<td>7.1</td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td></td>
<td><strong>61.4</strong></td>
<td>15.2</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>60.6</td>
<td>18.0</td>
</tr>
<tr>
<td>185 Rare Classes</td>
<td>Tighe et al CVPR13</td>
<td>19.0</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td><strong>26.4</strong></td>
<td><strong>14.4</strong></td>
</tr>
</tbody>
</table>
Results on LabelMe + SUN (LMSun)

- 45176 training images and 500 test images and the size of images ranges from 256x256 pixels to 800x600 pixels in 232 semantic labels.

- 185 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tighe et al IJCV12</td>
<td>54.9</td>
<td>7.1</td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>61.4</td>
<td>15.2</td>
</tr>
<tr>
<td>Ours</td>
<td>60.6</td>
<td>18.0</td>
</tr>
<tr>
<td>185 Rare Classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>185 Rare Classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>19.0</td>
<td>12.9</td>
</tr>
<tr>
<td>Ours</td>
<td>26.4</td>
<td>14.4</td>
</tr>
</tbody>
</table>
Results on LabelMe + SUN (LMSun)

- 45176 training images and 500 test images and the size of images ranges from 256x256 pixels to 800x600 pixels in 232 semantic labels.

- 185 rare classes from 80-20 rule.

<table>
<thead>
<tr>
<th>Accuracies</th>
<th>Per-Pixel</th>
<th>Per-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tighe et al IJCV12</td>
<td>54.9</td>
<td>7.1</td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>61.4</td>
<td>15.2</td>
</tr>
<tr>
<td>Ours</td>
<td>60.6</td>
<td>18.0</td>
</tr>
<tr>
<td>185 Rare Classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tighe et al CVPR13</td>
<td>19.0</td>
<td>12.9</td>
</tr>
<tr>
<td>Ours</td>
<td>26.4</td>
<td>14.4</td>
</tr>
</tbody>
</table>
Results on LabelMe + SUN (LMSun)
Results on LabelMe + SUN (LMSun)
Results on LabelMe + SUN (LMSun)
Concluding Remarks

- Handle rare classes from the long tail distribution.
- Introduce feedback based context filtering.
- Improve per-class accuracy without losing per-pixel accuracy.
Rare Class Expansion

Retrieval Set

Label distribution before expansion

Label distribution after expansion

Rare class expansion
Defining Rare Classes by Pareto Principle

The Pareto principle
The law of the vital few
Fundamental in visual recognition*

*X. Zhu, et al. CVPR14
Expansion Noise and Context Filtering

Irrelevant rare classes field, road and river become a new source of noise.
Observations in Image Retrieval

Retrieval set with labeled superpixels

Label distribution in retrieval set
Insufficient Samples

Retrieval set with labeled superpixels

Label distribution in retrieval set

Insufficient samples: Boat and person
Out-Of-Context Noise

Retrieval set with labeled superpixels

Label distribution in retrieval set

Outlier: building
Problems

Retrieval set with labeled superpixels

Sample insufficiency
Data noise

Label distribution in retrieval set