Scene Classification of Images and Video via Semantic Segmentation

by

Heather Dunlop

Digitalsmiths Corporation

Workshop on Perceptual Organization in Computer Vision
CVPR

June 13, 2010
Goal

- Identify scene types in video

mountain

beach

indoor

urban
Challenges

- It's not as easy as it sounds...
  - Viewpoint
  - Lighting
- Spatial arrangement
- Close-up shots
Related Work

- Lots of prior work on images:
  - Lazebnik, et al. (2006) Spatial pyramid matching
  - Oliva & Torralba (2001) Spatial envelope
Algorithm Overview

- Segment video into shots and scenes
- Select key frames
- On each key frame:
  - Classify scene as indoor or outdoor
  - If outdoor:
    - Semantic segmentation
    - Classify outdoor scene type with spatial pyramid
- Aggregate results across shots and scenes
Outline

- Semantic segmentation
- Scene classification of images
- Scene classification of video
Semantic Segmentation

- Goal: predict a material label for each image pixel
Segmentation

- Generate multiple segmentations
Feature Extraction

• For each segment, extract:
  • Color histogram
  • Edge strength and direction histograms
  • Line length histogram
  • Texton histogram
  • Shape metrics
Segment Merging

- Compute feature vector for each segment
- Compute difference of feature vectors for each adjacent pair
- Using Random Forest classifier, merge those most likely to belong to same material class
Material Classification

- Extract features for each region
- SVM for material classification
Semantic Segmentation Result

- Merge results across multiple segmentations

- building
- trees
- sky
- water
Scene Classification

- Goal:
  - Indoor
  - Outdoor
  - Undetermined
Indoor/Outdoor Classification

indoor
outdoor
undetermined
SVM

color, edge, line, texture features

indoor/outdoor/undetermined
SVM

indoor
outdoor
undetermined
SVM

indoor
outdoor
undetermined
SVM
Outdoor Classification

- Semantic segmentation
- Spatial pyramid
- SVMs for multi-label classification
Video

- Goal: extract scene types from a sequence of frames

open water  urban
Segmenting Video

From Frames to Shots to Scenes

key frames

classified key frames

average across shot

95th percentile across scene

open water   urban
Experiments

• Image data set:
  • LabelMe, Google Images, movie frames
  • For semantic segmentation: 1019 images
  • For scene classification: 9855 images

• Video data set:
  • 281 videos from 49 TV shows and 6 movies
    (110 hours of content)
  • Each scene labeled
Sample Results

highway

snow

building  person  road/sidewalk  rock  sand/gravel  sky/clouds  trees/bushes  water  miscellaneous

grass  snow/ice  vehicle  miscellaneous
Sample Results

open water

urban

building
grass
road/sidewalk
rock
sand/gravel
sky/clouds
snow/ice
trees/bushes
vehicle
water
miscellaneous
Sample Results

indoor

undetermined
# Semantic Segmentation

<table>
<thead>
<tr>
<th>True Class</th>
<th>building</th>
<th>grass</th>
<th>person</th>
<th>road/sidewalk</th>
<th>rock</th>
<th>sand/gravel</th>
<th>sky/clouds</th>
<th>snow/ice</th>
<th>trees/bushes</th>
<th>vehicle</th>
<th>water</th>
<th>miscellaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>building</td>
<td>71</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>grass</td>
<td>1</td>
<td>71</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>person</td>
<td>5</td>
<td>1</td>
<td>75</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>road/sidewalk</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>52</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>rock</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>47</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>sand/gravel</td>
<td>4</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>14</td>
<td>48</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>sky/clouds</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>91</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>snow/ice</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>13</td>
<td>41</td>
<td>0</td>
<td>2</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>trees/bushes</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>77</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>vehicle</td>
<td>22</td>
<td>0</td>
<td>15</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>51</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>water</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>11</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>66</td>
<td>1</td>
</tr>
<tr>
<td>miscellaneous</td>
<td>13</td>
<td>6</td>
<td>26</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>13</td>
<td>1</td>
<td>23</td>
</tr>
</tbody>
</table>
Scene Classification on Photographs

<table>
<thead>
<tr>
<th>Scene Type</th>
<th>Our Method</th>
<th>Lazebnik et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast/beach</td>
<td>.60</td>
<td>.44</td>
</tr>
<tr>
<td>Desert</td>
<td>.76</td>
<td>.48</td>
</tr>
<tr>
<td>Forest</td>
<td>.71</td>
<td>.84</td>
</tr>
<tr>
<td>Grassland</td>
<td>.79</td>
<td>.56</td>
</tr>
<tr>
<td>Highway</td>
<td>.67</td>
<td>.79</td>
</tr>
<tr>
<td>Lake/River</td>
<td>.44</td>
<td>.42</td>
</tr>
<tr>
<td>Mountainous</td>
<td>.73</td>
<td>.81</td>
</tr>
<tr>
<td>Open Water</td>
<td>.70</td>
<td>.67</td>
</tr>
<tr>
<td>Sky</td>
<td>.82</td>
<td>.83</td>
</tr>
<tr>
<td>Snow</td>
<td>.75</td>
<td>.69</td>
</tr>
<tr>
<td>Urban</td>
<td>.90</td>
<td>.87</td>
</tr>
<tr>
<td>Outdoor</td>
<td>.94</td>
<td>.99</td>
</tr>
<tr>
<td>Indoor</td>
<td>.73</td>
<td>.87</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>.73</strong></td>
<td><strong>.71</strong></td>
</tr>
</tbody>
</table>

Up to 28% per-category improvement.
### Scene Classification on Video

<table>
<thead>
<tr>
<th>Scene</th>
<th>Keyframes</th>
<th>Scenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast/beach</td>
<td>.13</td>
<td>.34</td>
</tr>
<tr>
<td>Desert</td>
<td>.04</td>
<td>.09</td>
</tr>
<tr>
<td>Forest</td>
<td>.29</td>
<td>.45</td>
</tr>
<tr>
<td>Grassland</td>
<td>.32</td>
<td>.47</td>
</tr>
<tr>
<td>Highway</td>
<td>.16</td>
<td>.33</td>
</tr>
<tr>
<td>Lake/River</td>
<td>.02</td>
<td>.07</td>
</tr>
<tr>
<td>Mountainous</td>
<td>.05</td>
<td>.11</td>
</tr>
<tr>
<td>Open Water</td>
<td>.33</td>
<td>.52</td>
</tr>
<tr>
<td>Sky</td>
<td>.24</td>
<td>.34</td>
</tr>
<tr>
<td>Snow</td>
<td>.04</td>
<td>.08</td>
</tr>
<tr>
<td>Urban</td>
<td>.33</td>
<td>.62</td>
</tr>
<tr>
<td>Outdoor</td>
<td>.67</td>
<td>.86</td>
</tr>
<tr>
<td>Indoor</td>
<td>.72</td>
<td>.82</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>.26</strong></td>
<td><strong>.39</strong></td>
</tr>
</tbody>
</table>

Method for aggregating across shots and scenes produces 13% improvement.
Conclusions

- We have developed a system that integrates:
  - segmentation
  - recognition of scene components
  - classification of whole images and video sequences
- Techniques that address the unique properties of video are a necessity
Future Work

- Incorporating face and body detection to identify when background is obstructed
- Background/motion segmentation
- Bag of features techniques for classifying material concepts
Thank you!

Questions?