Combining Appearance and Structure from Motion Features for Road Scene Understanding

Paul Sturgess, Karteek Alahari, Ľubor Ladický, Phil Torr
Oxford Brookes University

http://cms.brookes.ac.uk/research/visiongroup/
Goal: Classify ↔ Segment

- Abundance of street level imagery
- Classify every pixel in an image

The Cambridge-driving Labeled Video Database

http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/

Method

- A complementary set of features
  - Can describe a wide variety of object-classes

- Higher Order CRF
  - Produces high quality object-class boundaries

- Joint Boost for Unary Potentials
  - Single classifier for all features

- Evaluation
  - High quality annotated ground truth
Features

- Structure-from-motion
  - Moving Vs Static, 3D location cues, Texture

Features

• HOG
• Colour
• Location
• Textons
Higher Order CRF

\[ E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{(i,j) \in E} \psi_{ij}(x_i, x_j) + \sum_{c \in S} \psi_c(x_c) \]

- Likelihood of a pixel taking a label
- Computed via a boosting approach
Boosting for Unary Potentials

- **TextonBoost**
  - Context exploited
  - Boosted combination of textons
  - Response defined by the pair
    \[ \text{[texton } t, \text{ rectangular region } r] . \]

*TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation.* ECCV 2006.
Boosting for Unary Potentials

• Dense Boost
  • Response defined by the triplet
    \([\text{feature type } f, \text{ feature cluster } t, \text{ rectangular region } r]\)
    \(f = \{\text{SfM, HOG, Colour, Location, Texton}\}\)

Associative hierarchical crfs for object class image segmentation. ICCV 2009.
## Unary Potential Result

<table>
<thead>
<tr>
<th>Ground</th>
<th>Raw</th>
<th><strong>Road</strong></th>
<th><strong>Building</strong></th>
<th><strong>Sky</strong></th>
<th><strong>Tree</strong></th>
<th><strong>Sidewalk</strong></th>
<th><strong>Car</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Brostow</td>
<td></td>
<td>Building</td>
<td>Tree</td>
<td>Sky</td>
<td>Car</td>
<td>Sign</td>
<td>Road</td>
</tr>
<tr>
<td>46.2</td>
<td>61.9</td>
<td>89.7</td>
<td>68.6</td>
<td>42.9</td>
<td>89.5</td>
<td>53.6</td>
<td>46.6</td>
</tr>
<tr>
<td>Unary</td>
<td></td>
<td>61.9</td>
<td>67.3</td>
<td>91.1</td>
<td>71.1</td>
<td>58.5</td>
<td><strong>92.9</strong></td>
</tr>
</tbody>
</table>

Columns = Per-class recall, Average = Average recall, Global = Overall correctly labelled pixels
Higher Order CRF

\[ E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{(i,j) \in E} \psi_{ij}(x_i, x_j) + \sum_{c \in S} \psi_c(x_c) \]

- Contrast sensitive Potts model
- Encourages label consistency in adjacent pixels
Higher Order CRF

\[ E(x) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i, j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathcal{S}} \psi_c(x_c) \]

- Contrast sensitive Potts model
- Encourages label consistency in adjacent pixels
Pairwise Potential Result

<table>
<thead>
<tr>
<th></th>
<th>Building</th>
<th>Tree</th>
<th>Sky</th>
<th>Car</th>
<th>Sign</th>
<th>Road</th>
<th>Pedestrian</th>
<th>Fence</th>
<th>Column</th>
<th>Sidewalk</th>
<th>Bicyclist</th>
<th>Average</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brostow</td>
<td>46.2</td>
<td>61.9</td>
<td>89.7</td>
<td>68.6</td>
<td>42.9</td>
<td>89.5</td>
<td>53.6</td>
<td>46.6</td>
<td>0.7</td>
<td>60.5</td>
<td>22.5</td>
<td>53</td>
<td>69.1</td>
</tr>
<tr>
<td>Unary</td>
<td>61.9</td>
<td>67.3</td>
<td>91.1</td>
<td>71.1</td>
<td>58.5</td>
<td>92.9</td>
<td>49.5</td>
<td>37.6</td>
<td>25.8</td>
<td>77.8</td>
<td>24.7</td>
<td>59.8</td>
<td>76.4</td>
</tr>
<tr>
<td>+Pairwise</td>
<td><strong>70.7</strong></td>
<td><strong>70.8</strong></td>
<td><strong>94.7</strong></td>
<td><strong>74.4</strong></td>
<td>55.9</td>
<td><strong>94.1</strong></td>
<td>45.7</td>
<td>37.2</td>
<td>13</td>
<td><strong>79.3</strong></td>
<td>23.1</td>
<td><strong>59.9</strong></td>
<td><strong>79.8</strong></td>
</tr>
</tbody>
</table>

Columns = Per-class recall, Average = Average recall, Global = Overall correctly labelled pixels
Higher Order CRF

\[ E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathcal{C}} \psi_c(x_c) \]

- Potential takes the form of a robust \( P^N \) model
- Encourages label consistency within a super-pixel
- Super-pixels computed using meanshift

Robust $\mathbf{P}^N$ model

\[ \psi_c(x_c) = \begin{cases} 
\frac{N_i(x_c)}{Q} \gamma_{\max} & \text{if } N_i(x_c) \leq Q \\
\gamma_{\max} & \text{otherwise,} 
\end{cases} \]

Number of inconsistent pixels

Ensures cost of breaking a good segment is higher than that of a bad segment

Robust $\mathbf{P}^N$ code: [http://sots.brookes.ac.uk/lubor/](http://sots.brookes.ac.uk/lubor/)
Segment Quality

• Label inconsistency cost depends on segment quality

\[ \gamma_{\text{max}} = |c| \theta^\alpha \left( \theta^h_p + \theta^h_v G(c) \right) \]

• Low variance indicates good quality
• High variance indicates poor quality
Multiple Segmentations

• Single Segmentation?

• Combine multiple segmentations
HO Potential Result

<table>
<thead>
<tr>
<th></th>
<th>Road Void</th>
<th>Building Column</th>
<th>Sky Sign</th>
<th>Tree Fence</th>
<th>Sidewalk Pedestrian</th>
<th>Car Cyclist</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Brostow</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building</td>
<td>46.2</td>
<td>61.9</td>
<td>89.7</td>
<td>68.6</td>
<td>42.9</td>
<td>89.5</td>
</tr>
<tr>
<td>Tree</td>
<td>61.9</td>
<td>67.3</td>
<td>91.1</td>
<td>71.1</td>
<td><strong>58.5</strong></td>
<td>92.9</td>
</tr>
<tr>
<td>Sky</td>
<td>89.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>49.5</td>
</tr>
<tr>
<td>Car</td>
<td>68.6</td>
<td></td>
<td>42.9</td>
<td></td>
<td>89.5</td>
<td><strong>53.6</strong></td>
</tr>
<tr>
<td>Sign</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>46.6</td>
</tr>
<tr>
<td><strong>Road</strong></td>
<td>46.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>53.6</strong></td>
</tr>
<tr>
<td><strong>Pedestrian</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>Fence</td>
<td>37.6</td>
<td></td>
<td></td>
<td></td>
<td>25.8</td>
<td><strong>79.3</strong></td>
</tr>
<tr>
<td>Column</td>
<td>22.5</td>
<td></td>
<td>77.8</td>
<td>24.7</td>
<td>59.8</td>
<td><strong>79.3</strong></td>
</tr>
<tr>
<td>Sidewalk</td>
<td>53</td>
<td>22.5</td>
<td>76.4</td>
<td></td>
<td></td>
<td><strong>79.3</strong></td>
</tr>
<tr>
<td>Pedestrian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>28.5</strong></td>
</tr>
<tr>
<td>Bicyclist</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>59.2</td>
</tr>
<tr>
<td>Average</td>
<td>69.1</td>
<td>59.8</td>
<td>76.4</td>
<td></td>
<td></td>
<td><strong>83.8</strong></td>
</tr>
<tr>
<td>Global</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>83.8</strong></td>
</tr>
</tbody>
</table>

Columns = Per-class recall, Average = Average recall, Global = Overall correctly labelled pixels
Brostow et al
ECCV 08

Unary

+Pairwise

+HO

Ground Truth

Raw

<table>
<thead>
<tr>
<th>Road</th>
<th>Building</th>
<th>Sky</th>
<th>Tree</th>
<th>Sidewalk</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Void</td>
<td>Column</td>
<td>Sign</td>
<td>Fence</td>
<td>Pedestrian</td>
<td>Cyclist</td>
</tr>
</tbody>
</table>
Brostow et al
ECCV 08

Unary

+Pairwise

+HO

Ground Truth

Raw

<table>
<thead>
<tr>
<th>Road</th>
<th>Building</th>
<th>Sky</th>
<th>Tree</th>
<th>Sidewalk</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Void</td>
<td>Column</td>
<td>Sign</td>
<td>Fence</td>
<td>Pedestrian</td>
<td>Cyclist</td>
</tr>
</tbody>
</table>
Brostow et al
ECCV 08

Unary

+Pairwise

+HO

Ground Truth

Raw
HO Problems
## Evaluation Summary

<table>
<thead>
<tr>
<th></th>
<th>Building</th>
<th>Tree</th>
<th>Sky</th>
<th>Car</th>
<th>Sign-Symbol</th>
<th>Road</th>
<th>Pedestrian</th>
<th>Fence</th>
<th>Column-Pole</th>
<th>Sidewalk</th>
<th>Bicyclist</th>
<th>Average</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mot. [8]</td>
<td>43.9</td>
<td>46.2</td>
<td>79.5</td>
<td>44.6</td>
<td>19.5</td>
<td>82.5</td>
<td>24.4</td>
<td>58.8</td>
<td>0.1</td>
<td>61.8</td>
<td>18.0</td>
<td>43.6</td>
<td>61.8</td>
</tr>
<tr>
<td>App. [8]</td>
<td>38.7</td>
<td>60.7</td>
<td>90.1</td>
<td>71.1</td>
<td>51.4</td>
<td>88.6</td>
<td>54.6</td>
<td>40.1</td>
<td>1.1</td>
<td>55.5</td>
<td>23.6</td>
<td>52.3</td>
<td>66.5</td>
</tr>
<tr>
<td>Combined [8]</td>
<td>46.2</td>
<td>61.9</td>
<td>89.7</td>
<td>68.6</td>
<td>42.9</td>
<td>89.5</td>
<td>53.6</td>
<td>46.6</td>
<td>0.7</td>
<td>60.5</td>
<td>22.5</td>
<td>53.0</td>
<td>69.1</td>
</tr>
<tr>
<td>$\psi_i$</td>
<td>61.9</td>
<td>67.3</td>
<td>91.1</td>
<td>71.1</td>
<td>58.5</td>
<td>92.9</td>
<td>49.5</td>
<td>37.6</td>
<td>25.8</td>
<td>77.8</td>
<td>24.7</td>
<td>59.8</td>
<td>76.4</td>
</tr>
<tr>
<td>$\psi_i + \psi_{ij}$</td>
<td>70.7</td>
<td>70.8</td>
<td>94.7</td>
<td>74.4</td>
<td>55.9</td>
<td>94.1</td>
<td>45.7</td>
<td>37.2</td>
<td>13.0</td>
<td>79.3</td>
<td>23.1</td>
<td>59.9</td>
<td>79.8</td>
</tr>
<tr>
<td>$\psi_i + \psi_{ij} + \psi_c$</td>
<td><strong>84.5</strong></td>
<td><strong>72.6</strong></td>
<td><strong>97.5</strong></td>
<td>72.7</td>
<td>34.1</td>
<td><strong>95.3</strong></td>
<td>34.2</td>
<td>45.7</td>
<td>8.1</td>
<td>77.6</td>
<td><strong>28.5</strong></td>
<td>59.2</td>
<td><strong>83.8</strong></td>
</tr>
</tbody>
</table>

Columns = Per-class recall, Average = Average recall, Global = Overall correctly labelled pixels

- Improvement in 9 out of 11 classes
- Pairwise terms improve most classes
- Higher order terms further improve most classes
## Evaluation Summary

<table>
<thead>
<tr>
<th></th>
<th>Building</th>
<th>Tree</th>
<th>Sky</th>
<th>Car</th>
<th>Sign-Symbol</th>
<th>Road</th>
<th>Pedestrian</th>
<th>Fence</th>
<th>Column-Pole</th>
<th>Sidewalk</th>
<th>Bicyclist</th>
<th>Average</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mot. [8]</td>
<td>43.9</td>
<td>46.2</td>
<td>79.5</td>
<td>44.6</td>
<td>19.5</td>
<td>82.5</td>
<td>24.4</td>
<td><strong>58.8</strong></td>
<td>0.1</td>
<td>61.8</td>
<td>18.0</td>
<td>43.6</td>
<td>61.8</td>
</tr>
<tr>
<td>App. [8]</td>
<td>38.7</td>
<td>60.7</td>
<td>90.1</td>
<td>71.1</td>
<td>51.4</td>
<td>88.6</td>
<td><strong>54.6</strong></td>
<td>40.1</td>
<td>1.1</td>
<td>55.5</td>
<td>23.6</td>
<td>52.3</td>
<td>66.5</td>
</tr>
<tr>
<td>Combined [8]</td>
<td>46.2</td>
<td>61.9</td>
<td>89.7</td>
<td>68.6</td>
<td>42.9</td>
<td>89.5</td>
<td>53.6</td>
<td>46.6</td>
<td>0.7</td>
<td>60.5</td>
<td>22.5</td>
<td>53.0</td>
<td>69.1</td>
</tr>
<tr>
<td>$\psi_i$</td>
<td>61.9</td>
<td>67.3</td>
<td>91.1</td>
<td>71.1</td>
<td><strong>58.5</strong></td>
<td>92.9</td>
<td>49.5</td>
<td>37.6</td>
<td><strong>25.8</strong></td>
<td>77.8</td>
<td>24.7</td>
<td>59.8</td>
<td>76.4</td>
</tr>
<tr>
<td>$\psi_i + \psi_{ij}$</td>
<td>70.7</td>
<td>70.8</td>
<td>94.7</td>
<td><strong>74.4</strong></td>
<td>55.9</td>
<td>94.1</td>
<td>45.7</td>
<td>37.2</td>
<td>13.0</td>
<td><strong>79.3</strong></td>
<td>23.1</td>
<td><strong>59.9</strong></td>
<td>79.8</td>
</tr>
<tr>
<td>$\psi_i + \psi_{ij} + \psi_c$</td>
<td><strong>84.5</strong></td>
<td><strong>72.6</strong></td>
<td><strong>97.5</strong></td>
<td>72.7</td>
<td>34.1</td>
<td><strong>95.3</strong></td>
<td>34.2</td>
<td>45.7</td>
<td>8.1</td>
<td>77.6</td>
<td><strong>28.5</strong></td>
<td>59.2</td>
<td><strong>83.8</strong></td>
</tr>
</tbody>
</table>

Columns = Per-class recall, Average = Average recall, Global = Overall correctly labelled pixels

- Improvement in 9 out of 11 classes
- Pairwise terms improve most classes
- Higher order terms further improve most classes
- Brostow et al ECCV08 better for 2 classes
• Column/pole = 2,536,704 << building = 57,583,181
• Poorer on all classes below 2% training pixels
### Discussion: HO Problems

<table>
<thead>
<tr>
<th></th>
<th>Building</th>
<th>Tree</th>
<th>Sky</th>
<th>Car</th>
<th>Sign-Symbol</th>
<th>Road</th>
<th>Pedestrian</th>
<th>Fence</th>
<th>Column-Pole</th>
<th>Sidewalk</th>
<th>Bicyclist</th>
<th>Average</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_i$</td>
<td>61.9</td>
<td>67.3</td>
<td>91.1</td>
<td>71.1</td>
<td>58.5</td>
<td>92.9</td>
<td>49.5</td>
<td>37.6</td>
<td>25.8</td>
<td>77.8</td>
<td>24.7</td>
<td>59.8</td>
<td>76.4</td>
</tr>
<tr>
<td>$\psi_i + \psi_{ij}$</td>
<td>70.7</td>
<td>70.8</td>
<td>94.7</td>
<td><strong>74.4</strong></td>
<td>55.9</td>
<td>94.1</td>
<td>45.7</td>
<td>37.2</td>
<td>13.0</td>
<td><strong>79.3</strong></td>
<td>23.1</td>
<td><strong>59.9</strong></td>
<td>79.8</td>
</tr>
<tr>
<td>$\psi_i + \psi_{ij} + \psi_c$</td>
<td><strong>84.5</strong></td>
<td><strong>72.6</strong></td>
<td><strong>97.5</strong></td>
<td>72.7</td>
<td>34.1</td>
<td><strong>95.3</strong></td>
<td>34.2</td>
<td>45.7</td>
<td>8.1</td>
<td>77.6</td>
<td><strong>28.5</strong></td>
<td>59.2</td>
<td><strong>83.8</strong></td>
</tr>
</tbody>
</table>

Columns = Per-class recall, Average = Average recall, Global = Overall correctly labelled pixels

- **Decrease doesn't match with qualitative results**
Discussion: Error

Recall = \frac{\text{Ground Truth}}{\text{Predicted}}

= 100\% \text{ for column/pole}

- Favours over estimates
Discussion: Error

Intersection/union =

Ground Truth

Predicted

Ground Truth

Predicted

= Almost 0% for column/pole

• Allows for an independent per-class error measurement
• Penalises both over- and under-estimates

Slide adapted from
### Discussion: Error

- Intersection/union table

<table>
<thead>
<tr>
<th></th>
<th>Building</th>
<th>Tree</th>
<th>Sky</th>
<th>Car</th>
<th>Sign-Symbol</th>
<th>Road</th>
<th>Pedestrian</th>
<th>Fence</th>
<th>Column-Pole</th>
<th>Sidewalk</th>
<th>Bicyclist</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_i$</td>
<td>55.3</td>
<td>54.3</td>
<td>84.8</td>
<td>51.8</td>
<td>11.9</td>
<td>85.5</td>
<td>15.6</td>
<td>27.4</td>
<td>7.5</td>
<td>60.0</td>
<td>15.7</td>
<td>42.71</td>
</tr>
<tr>
<td>$\psi_i + \psi_{ij}$</td>
<td>63.6</td>
<td>58.0</td>
<td>87.8</td>
<td>55.9</td>
<td>13.6</td>
<td>86.4</td>
<td>16.9</td>
<td>27.6</td>
<td>6.1</td>
<td>61.9</td>
<td>18.1</td>
<td>45.07</td>
</tr>
<tr>
<td>$\psi_i + \psi_{ij} + \psi_c$</td>
<td>71.6</td>
<td>60.4</td>
<td>89.5</td>
<td>58.3</td>
<td>19.4</td>
<td>86.6</td>
<td>26.1</td>
<td>35.0</td>
<td>7.2</td>
<td>63.8</td>
<td>22.6</td>
<td>49.15</td>
</tr>
</tbody>
</table>

- Higher Order terms improve performance in all classes
Conclusion

- Strong unary potential from boosting
- HO terms yield more precise boundaries
- Improvement in 9 out of 11 classes
- Intersection/union error more informative
- Directions
  - Balance training data
  - Potentials for thin structures
  - Use Associative hierarchical CRFs


Ass ociative hierarchical crfs for object class image segmentation. ICCV 2009.
Questions

Raw Image

Unary + Pairwise

Ground Truth

Unary + Pairwise + Higher Order

<table>
<thead>
<tr>
<th>Road</th>
<th>Building</th>
<th>Sky</th>
<th>Tree</th>
<th>Sidewalk</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Void</td>
<td>Column</td>
<td>Sign</td>
<td>Fence</td>
<td>Pedestrian</td>
<td>Cyclist</td>
</tr>
</tbody>
</table>