What, Where & How Many?
Combining Object Detectors and CRFs

Ľúbor Ladický, Paul Sturgess, Karteek Alahari,
Christopher Russell, Philip H.S. Torr

http://cms.brookes.ac.uk/research/visiongroup/
Complete Scene Understanding

Involves

- Localization of all instances of foreground objects ("things")
- Localization of all background classes ("stuff")
- Pixel-wise segmentation
- 3D reconstruction
- Pose detection
- Action recognition
- Event recognition
Complete Scene Understanding

Involves

- Localization of all instances of foreground objects (“things”)
- Localization of all background classes (“stuff”)
- Pixel-wise segmentation
- 3D reconstruction
- Pose detection
- Action recognition
- Event recognition

Semantic Segmentation
Complete Scene Understanding

Involves

- Localization of all instances of foreground objects ("things")
- Localization of all background classes ("stuff")
- Pixel-wise segmentation
Semantic Scene Understanding

We're interested in whole scene understanding
Given an image, detect every *thing* in it.

*Thing*: An object with a specific size and shape.

*Adelson, Forsyth et al. 96*
We're interested in whole scene understanding Given an image, detect every *thing* in it.

**Thing**: An object with a specific size and shape.

Adelson, Forsyth et al. 96
We're interested in whole scene understanding. Given an image, detect every *thing* in it.

**Thing**: An object with a specific size and shape.

*Adelson, Forsyth et al. 96*
Semantic Scene Understanding

We're interested in whole scene understanding. Given an image, detect every *thing* in it.

**Thing**: An object with a specific size and shape.
We're interested in whole scene understanding. Given an image, detect every *thing* in it.

**Thing**: An object with a specific size and shape.

*Adelson, Forsyth et al. 96*
We're interested in whole scene understanding
Given an image, label all the stuff

Stuff: Material defined by a homogeneous or repetitive pattern, with no specific spatial extent / shape.

Adelson, Forsyth et al. 96
Semantic Scene Understanding

We're interested in whole scene understanding
Given an image, label all the *stuff*

*Stuff* : Material defined by a homogeneous or repetitive pattern, with no specific spatial extent / shape.

Adelson, Forsyth et al. 96
Our plan is to combine

- State of the art sliding window object detection

- State of the art segmentation techniques
Algorithms for Object Localization

- Sliding window detectors
  - HOG descriptor (Dalal & Triggs CVPR05)
  - Based on histograms of features (Vedaldi et al. ICCV09)
  - Part-based models (Felzenszwalb et al. CVPR09)
Algorithms for Object Localization

Sliding window detectors

- HOG descriptor (Dalal & Triggs CVPR05)
- Based on histograms of features (Vedaldi et al. ICCV09)
- Part-based models (Felzenszwalb et al. CVPR09)
Algorithms for Object Localization

- Sliding window detectors
  - HOG descriptor (Dalal & Triggs CVPR05)
  - Based on histograms of features (Vedaldi et al. ICCV09)
  - Part-based models (Felzenszwalb et al. CVPR09)
Non-maxima suppression

- Greedily
- Using field of indicator variables (Ramanan et al ICCV09)
  - One binary variable $y_i \in \{0, 1\}$ per location/scale
Non-maxima suppression

- Greedily
- Using field of indicator variables (Ramanan et al ICCV09)
  - One binary variable $y_i \in \{0, 1\}$ per location/scale
Sliding window detectors

- Sliding window + Segmentation
  - OBJCUT (Kumar et al. 05)
  - Updating colour model (GrabCut - Rother et al. 04)
Sliding window detectors not good for “stuff”

Try to detect “sky”!
Sliding window detectors not good for “stuff”

Sky is irregular shape not suited to the sliding window approach
Sliding window detectors

- State-of-the-art for object detection ("things")
- Do not work for background classes ("stuff")
  - No distinct shape
  - Cannot be enclosed in a box
- Cannot recover from incorrect detections
Pairwise CRF over pixels

Input image

Final segmentation

Shotton et al. ECCV06
Pairwise CRF over pixels

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

Shotton et al. ECCV06
Pairwise CRF over pixels

Input image

Final segmentation

Training of Potentials

CRF construction

$E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{i \in V, j \in N_i} \psi_{ij}(x_i, x_j)$

Shotton et al. ECCV06
Pairwise CRF over Super-pixels / Segments

Input image

Unsupervised segmentation

Algorithms for Object-class Segmentation

Pairwise CRF over Super-pixels / Segments

Input image

Unsupervised segmentation

Algorithms for Object-class Segmentation

Pairwise CRF over Super-pixels / Segments

Unsupervised segmentation

Input image

Training of potentials

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

Batra et al. CVPR08, Yang et al. CVPR07, Zitnick et al. CVPR08, Rabinovich et al. ICCV07, Boix et al. CVPR10
Pairwise CRF over Super-pixels / Segments

Unsupervised segmentation

Training of potentials

MAP

$E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{i \in V, j \in N_i} \psi_{ij}(x_i, x_j)$

Batra et al. CVPR08, Yang et al. CVPR07, Zitnick et al. CVPR08, Rabinovich et al. ICCV07, Boix et al. CVPR10
Associative Hierarchical CRF

Input image

Multiple segmentations or hierarchies

Ladický et al. ICCV09
Algorithms for Object-class Segmentation

Associative Hierarchical CRF

Input image

Multiple segmentations or hierarchies

CRF construction

Ladický et al. ICCV09
Algorithms for Object-class Segmentation

Associative Hierarchical CRF

Input image

Multiple segmentations or hierarchies

CRF construction

Final segmentation

Ladický et al. ICCV09, Russell et al. UAI10
• State-of-the-art performance on MSRC & CamVid
• Merges information at multiple scales
• Can recover from incorrect segmentations
Associative Hierarchical CRF

- State-of-the-art performance on MSRC & CamVid
- Merges information at multiple scales
- Can recover from incorrect segmentations

However,
- No concept of “things”
- Cannot distinguish between multiple instances

Ladický et al. ICCV09, Sturgess et al., BMVC09
CRF Formulation with Detectors

- CRF formulation altered with a potential for each detection
CRF Formulation with Detectors

- CRF formulation altered with a potential for each detection

\[
E(x) = E_{pix}(x) + \sum_{d \in D} \psi_d(x_d, H_d, l_d)
\]

AH-CRF energy without detectors

CRF graph over pixels
CRF Formulation with Detectors

- CRF formulation altered with a potential for each detection

\[ E(x) = E_{pix}(x) + \sum_{d \in D} \psi_d(x_d, H_d, l_d) \]

AH-CRF energy without detectors

Set of pixels of d-th detection

Classifier response

Detected label

CRF graph over pixels
Joint CRF formulation should contain:
- Possibility to reject detection hypothesis
- Recover the status of the detection \((0 / 1)\)

Thus, potential is a minimum over indicator variable \(y_d \in \{ 0, 1 \}\)

\[
\psi_d(x_d, H_d, l_d) = \min_{y_d} \phi_d(y_d, x_d, H_d, l_d)
\]
CRF Formulation with Detectors

- Detection potential should decrease the energy of labelling agreeing with the detection hypothesis
- Partial disagreement penalized but not directly rejects the hypothesis
CRF Formulation with Detectors

- Detection potential should decrease the energy of labelling agreeing with the detection hypothesis.
- Partial disagreement penalized but not directly rejects the hypothesis.

\[
\psi_d(x_d, H_d, l_d) = \min_{y_d} (-f(x_d, H_d)y_d + g(N_d, H_d)y_d)
\]

Detection hypothesis status \rightarrow Detection strength \rightarrow Partial inconsistency cost

CRF graph over pixels
CRF Formulation with Detectors

- Detection potential should decrease the energy of labelling agreeing with the detection hypothesis.
- Partial disagreement penalized but not directly rejects the hypothesis.

\[
\psi_d(x_d, H_d, l_d) = \min_{y_d} \left( -f(x_d, H_d)y_d + g(N_d, H_d)y_d \right)
\]

Linear thresholded

![Diagram showing CRF graph over pixels with detections and pixels highlighted.](image-url)
This higher order potential can be transformed to

\[
\psi_d(x_d, H_d) = \min_{y_d \in \{0,1\}} (-f(x_d, H_d)y_d + k_d N_d y_d)
\]

\[
= -f(x_d, H_d) + \min(f(x_d, H_d), k_d \sum_{j \in x_d} \delta(x_j \neq l_d))
\]

which take the form of Robust P^N (Kohli et al. CVPR08)

\[
\psi_h(x) = \min(\gamma_{max}, \min_l (\gamma_l + k_l \sum_{i \in x} \delta(x_i \neq l)))
\]

Solvable with all graph cut-based methods
Results on CamVid dataset

Brostow et al. ECCV08, Sturgess et al. BMVC09
Results on CamVid dataset

Result without detections

Set of detections

Final Result

Also provides number of object instances (using $y_d$’s)
## Results on VOC2009 dataset

<table>
<thead>
<tr>
<th>Input image</th>
<th>CRF without detectors</th>
<th>CRF with detectors</th>
<th>Input image</th>
<th>CRF without detectors</th>
<th>CRF with detectors</th>
<th>CRF with detectors</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Input image" /></td>
<td><img src="image2.png" alt="CRF without detectors" /></td>
<td><img src="image3.png" alt="CRF with detectors" /></td>
<td><img src="image4.png" alt="Input image" /></td>
<td><img src="image5.png" alt="CRF without detectors" /></td>
<td><img src="image6.png" alt="CRF with detectors" /></td>
<td><img src="image7.png" alt="CRF with detectors" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Background</th>
<th>Aeroplane</th>
<th>Bicycle</th>
<th>Bird</th>
<th>Boat</th>
<th>Bottle</th>
<th>Bus</th>
<th>Car</th>
<th>Cat</th>
<th>Chair</th>
<th>Cow</th>
<th>Dining table</th>
<th>Dog</th>
<th>Horse</th>
<th>Motor bike</th>
<th>Person</th>
<th>Potted plant</th>
<th>Sheep</th>
<th>Sofa</th>
<th>Train</th>
<th>TV/monitor</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>BONN_SVM-SEGME</td>
<td>83.9</td>
<td>64.3</td>
<td>21.8</td>
<td>21.7</td>
<td>32.0</td>
<td>40.2</td>
<td>57.3</td>
<td>49.4</td>
<td>38.8</td>
<td>5.2</td>
<td>28.5</td>
<td>22.0</td>
<td>19.6</td>
<td>33.6</td>
<td>45.5</td>
<td>33.6</td>
<td>27.3</td>
<td>40.4</td>
<td>18.1</td>
<td>33.6</td>
<td>46.1</td>
<td>36.3</td>
</tr>
<tr>
<td>CVC_HOCRF</td>
<td>80.2</td>
<td>67.1</td>
<td>26.6</td>
<td>30.3</td>
<td>31.6</td>
<td>30.0</td>
<td>44.5</td>
<td>41.6</td>
<td>25.2</td>
<td>5.9</td>
<td>27.8</td>
<td>11.0</td>
<td>23.1</td>
<td>40.5</td>
<td>53.2</td>
<td>32.0</td>
<td>22.2</td>
<td>37.4</td>
<td>23.6</td>
<td>40.3</td>
<td>30.2</td>
<td>34.5</td>
</tr>
<tr>
<td>UOCITLI_SVM-MDPM</td>
<td>78.9</td>
<td>35.3</td>
<td>22.5</td>
<td>19.1</td>
<td>23.5</td>
<td>36.2</td>
<td>41.2</td>
<td>50.1</td>
<td>11.7</td>
<td>8.9</td>
<td>28.5</td>
<td>1.4</td>
<td>5.9</td>
<td>24.0</td>
<td>35.3</td>
<td>33.4</td>
<td>35.1</td>
<td>27.7</td>
<td>14.2</td>
<td>34.1</td>
<td>41.8</td>
<td>29.0</td>
</tr>
<tr>
<td>NECUIUC_CLS-DTCT</td>
<td>81.8</td>
<td>41.9</td>
<td>23.1</td>
<td>22.4</td>
<td>22.0</td>
<td>27.8</td>
<td>43.2</td>
<td>51.8</td>
<td>25.9</td>
<td>4.5</td>
<td>18.5</td>
<td>18.0</td>
<td>23.5</td>
<td>26.9</td>
<td>36.6</td>
<td>34.8</td>
<td>8.8</td>
<td>28.3</td>
<td>14.0</td>
<td>35.5</td>
<td>34.7</td>
<td>29.7</td>
</tr>
<tr>
<td>LEAR_SEGDGT</td>
<td>79.1</td>
<td>44.6</td>
<td>15.5</td>
<td>20.5</td>
<td>13.3</td>
<td>28.8</td>
<td>29.3</td>
<td>35.8</td>
<td>25.4</td>
<td>4.4</td>
<td>20.3</td>
<td>1.3</td>
<td>16.4</td>
<td>28.2</td>
<td>30.0</td>
<td>24.5</td>
<td>12.2</td>
<td>31.5</td>
<td>18.3</td>
<td>28.8</td>
<td>31.9</td>
<td>25.7</td>
</tr>
<tr>
<td>BROOKESMSRC_AHCRF</td>
<td>79.6</td>
<td>48.3</td>
<td>6.7</td>
<td>19.1</td>
<td>10.0</td>
<td>16.6</td>
<td>32.7</td>
<td>38.1</td>
<td>25.3</td>
<td>5.5</td>
<td>9.4</td>
<td>25.1</td>
<td>13.3</td>
<td>12.3</td>
<td>35.5</td>
<td>20.7</td>
<td>27.7</td>
<td>17.1</td>
<td>18.4</td>
<td>37.5</td>
<td>36.4</td>
<td>24.8</td>
</tr>
<tr>
<td>Our method</td>
<td>81.2</td>
<td>46.1</td>
<td>15.4</td>
<td>24.6</td>
<td>20.9</td>
<td>36.9</td>
<td>50.0</td>
<td>43.9</td>
<td>28.4</td>
<td><strong>11.5</strong></td>
<td>18.2</td>
<td><strong>25.4</strong></td>
<td>14.7</td>
<td>25.1</td>
<td>37.7</td>
<td>34.1</td>
<td>27.7</td>
<td>29.6</td>
<td>18.4</td>
<td><strong>43.8</strong></td>
<td>40.8</td>
<td>32.1</td>
</tr>
</tbody>
</table>
Summary

- Novel formulation of CRF with detectors
- Can recover from incorrect detections
- Possible to obtain instances of objects
- Efficiently solvable
Future and ongoing work

- Automatic non-maximum suppression
- Part-based models in CRF
- New things & stuff dataset available soon!
- Questions?