A Tour of Image Segmentation

Dr. Scott Cohen | Adobe Research
Image Segmentation

- Segment image pixels into different classes

- 2 Classes: Boy (Foreground), Not Boy (Background)
Scott Cohen, Gregg Wilensky, Jeff Chien (Adobe)
Segmentation Variations

- How many classes?
- How are the classes defined?
- What features are used to compute the segmentation?
- Hard Segmentation or Soft Segmentation?
  - Hard: a pixel is assigned to exactly one class
  - Soft: a pixel may be assigned to more than one class
- Automatic or Interactive computation?
  - What user input is provided?
- How many images are segmented?
Interactive Binary Segmentation

- User Strokes (Scribbles) in Foreground and Background

User Intent

Desired Segmentation

Interactive Graph Cuts for Optimal Boundary & Region Segmentation
Boykov, Jolly (ICCV 2001)
Interactive Binary Segmentation

- **GrabCut**: Draw a rectangle around the object to select

GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts
Rother, Kolmogorov, Blake (Siggraph 2004)
Interactive Binary Segmentation

- Magnetic Lasso: Trace around the object to select

Intelligent Scissors for Image Composition
Mortensen, Barrett (Siggraph 1995)
Automatic Binary Segmentation

- Segment the “Salient” Region
Automatic Binary Segmentation

- Segment the “Salient” Region
Automatic Binary Segmentation

- Segment the “Salient” Region
Automatic Binary Segmentation

- Segment In-focus Regions

Input Image

Output Segmentation
Image Matting

- Soft Binary Segmentation

\[ I_p = \alpha_p F_p + (1 - \alpha_p) B_p \]

Input Image I

Output Segmentation \( \alpha \in [0,1] \)
Image Matting: Compositing Application

\[ I_p = \alpha_p F_p + (1 - \alpha_p) B_p \]

\[ \hat{I}_p = \alpha_p^{01} H_p + (1 - \alpha_p^{01}) \hat{B}_p \]
Image Matting: Trimap Input

\[ I_p = \alpha_p F_p + (1 - \alpha_p) B_p \]

Unknown Region

Known Background

Known Foreground
Demo: Interactive Matting

- Brian Price, Scott Cohen (Adobe)
Co-Segmentation

- Segment the object in common in multiple images

Cosegmentation of Image Pairs by Histogram Matching
Rother, Kolmogorov, Minka, Blake (CVPR 2006)
## Co-Segmentation Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Foregrounds</th>
<th>Backgrounds</th>
<th>Automatic or Interactive?</th>
<th># of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram Matching (CVPR06)</td>
<td>Same</td>
<td>Different</td>
<td>Automatic</td>
<td>2</td>
</tr>
</tbody>
</table>
Co-Segmentation: Similar BGs, Interactive, Many Images

iCoseg: Interactive Co-segmentation with Intelligent Scribble Guidance
Batra, Kowdle, Parikh, Luo, Chen (CVPR 2010)
<table>
<thead>
<tr>
<th>Method</th>
<th>Foregrounds</th>
<th>Backgrounds</th>
<th>Automatic or Interactive?</th>
<th># of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram Matching (CVPR06)</td>
<td>Same</td>
<td>Different</td>
<td>Automatic</td>
<td>2</td>
</tr>
<tr>
<td>iCoseg (CVPR10)</td>
<td>Similar</td>
<td>Similar</td>
<td>Interactive</td>
<td>Many</td>
</tr>
</tbody>
</table>
Stereo Co-Segmentation: Same BGs, Interactive, 2 Images

StereoCut: Consistent Interactive Object Selection in Stereo Image Pairs
Price, Cohen (ICCV 2011)
Stereo Co-Segmentation Applications

- Localized Stereo Editing
- Stereo Inpainting: remove Co-Segmented object from the stereo picture

PatchMatch-based Content Completion of Stereo Image Pairs
Morse, Howard, Cohen, Price (3DimPVT 2012)
WHAT'S YOUR IDENTITY??

How do you define yourself by the clothes you wear and the products you buy? By the people you spend your time with? By the things you think and do?

SOUTH CAMDEN COMMUNITY SCHOOL

Ashley Smith Smith currently lives in the Camden area, with her. She works in marketing with a local company. She has been in the Sacramento Valley since 2005.

EAST SEATTLE ELEMENTARY SCHOOL

In 2002, Ashley Smith Smith was asked to teach in the East Seattle Elementary School. She was intrigued by the opportunity to work with children and help them learn.
## Co-Segmentation Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Foregrounds</th>
<th>Backgrounds</th>
<th>Automatic or Interactive?</th>
<th># of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram Matching (CVPR06)</td>
<td>Same</td>
<td>Different</td>
<td>Automatic</td>
<td>2</td>
</tr>
<tr>
<td>iCoseg (CVPR10)</td>
<td>Similar</td>
<td>Similar</td>
<td>Interactive</td>
<td>Many</td>
</tr>
<tr>
<td>Stereo (ICCV11)</td>
<td>Same</td>
<td>Same</td>
<td>Interactive</td>
<td>2</td>
</tr>
</tbody>
</table>
Co-Segmentation: Similar BGs, Automatic, Many Images

Unsupervised Joint Object Discovery and Segmentation in Internet Images
Rubinstein, Joulin, Kopf, Liu (CVPR 2013)

“Car” Internet Search

Input

Output
<table>
<thead>
<tr>
<th>Method</th>
<th>Foregrounds</th>
<th>Backgrounds</th>
<th>Automatic or Interactive?</th>
<th># of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram Matching (CVPR06)</td>
<td>Same</td>
<td>Different</td>
<td>Automatic</td>
<td>2</td>
</tr>
<tr>
<td>iCoseg (CVPR10)</td>
<td>Similar</td>
<td>Similar</td>
<td>Interactive</td>
<td>Many</td>
</tr>
<tr>
<td>Stereo (ICCV11)</td>
<td>Same</td>
<td>Same</td>
<td>Interactive</td>
<td>2</td>
</tr>
<tr>
<td>Object Discovery (CVPR13)</td>
<td>Similar, but more variation than iCoseg</td>
<td>Similar and Different</td>
<td>Automatic</td>
<td>Many</td>
</tr>
</tbody>
</table>
Scene Parsing | Semantic Segmentation

- Label each pixel in an image with its semantic category

Input Image

Desired Output

Legend:
- unlabeled
- building
- car
- person
- pole
- road
- sidewalk
- sign
- trash can
- tree
Details About Some Segmentation Methods

- **Common Framework**
- **Notation: Segmentation** \( X = \{ x_p \} \)

\[
I = \{ I_p \}
\]

\[
x_p = 1 \quad X = \{ x_p \} x_p = 0
\]
Common Energy Minimization Framework

- Energy function to measure quality of segmentation $X = \{x_p\}$

$$E(X) = \sum_p D_p(x_p) + \lambda \sum_{p,q \in N} V_{pq}(x_p, x_q)$$

* Global minimum found by min graphcut / maxflow algorithms

---

**Data Term**

- $D_p(0)$ = cost of labeling pixel $p$ as BG
- $D_p(1)$ = cost of labeling pixel $p$ as FG

**Smoothness Term**

- $V_{pq}(1,0)$ = cost of $p$ as FG, $q$ as BG
- $V_{pq}(0,1)$ = cost of $p$ as BG, $q$ as FG
- $V_{pq}(0,0)$ = cost of $p$ as BG, $q$ as BG = 0
- $V_{pq}(1,1)$ = cost of $p$ as FG, $q$ as FG = 0
Energy E for Stroke-based Binary Segmentation using Color

- K-means on FG Strokes

- K-means on BG Strokes

- Probability \( P_p(FG \mid I_p) \)

\[
D_p(FG) = \begin{cases} 
0 & \text{if } p \in F \\
\infty & \text{if } p \in B \\
- \log P_p(FG \mid I_p) & \text{otherwise}
\end{cases}
\]

\[
D_p(BG) = \begin{cases} 
0 & \text{if } p \in B \\
\infty & \text{if } p \in F \\
- \log P_p(BG \mid I_p) & \text{otherwise}
\end{cases}
\]
Energy $E$ for Binary Segmentation: Smoothness

$V_{pq}(BG, BG) = V_{pq}(FG, FG) = 0$

- Encourage segmentation boundaries to occur at image edges

$$V_{pq}(FG, BG) = \exp\left(-\frac{\|I_p - I_q\|^2}{2\sigma^2}\right)$$

$$V_{pq}(BG, FG) = V_{pq}(FG, BG)$$

$$\nabla I$$
Why is V called the smoothness term?

\[ E(X) = \sum_{p} D_p(x_p) + \lambda \sum_{p,q \in \mathbb{N}} V_{pq}(x_p, x_q) \]
In-Focus Segmentation: Frequency Decomposition
In-Focus Segmentation: Modeling Defocus Blur

- Larger blurs remove higher frequencies
Out-of-Focus Blur Estimation

- Measure power across frequencies
Automatic In-Focus Segmentation Results

\[ E(X) = \sum_p D_p(x_p) + \lambda \sum_{p,q \in N} V_{pq}(x_p, x_q) \]
Automatic In-Focus Segmentation Results

- Works when foreground and background colors are similar
Automatic In-Focus Segmentation Results
Interactive Stereo Co-Segmentation

Left

Right
Energy $E$ for Stereo Segmentation

- Add an energy term to include correspondence information

- If $(p_L, p_R)$ likely correspond, then their labels should be same

$$E(X) = \sum_p D_p(x_p) + \lambda \sum_{p,q \in N} V_{pq}(x_p, x_q) + \mu \sum_{p_L, p_R} P_{match}(p_L, p_R) |x_{p_L} - x_{p_R}|$$
Stereo Segmentation Results
Scene Parsing | Semantic Segmentation

- Label each pixel in an image with its semantic category

Input Image

Desired Output

Legend:
- unlabeled
- building
- car
- person
- pole
- road
- sidewalk
- sign
- trash can
- tree
Scene Parsing | Semantic Segmentation

- Training Data: Labeled Images (Input)
- SIFTflow: 2488 Labeled Training Images, 33 Classes
- LMSun: 45176 Labeled Training Images, 232 Classes
Scene Parsing | Semantic Segmentation

Input Image

Image Retrieval

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

\[ P_p(x_p = \text{sea}), P_p(x_p = \text{sand}), \ldots \]
Tighe, Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. CVPR13

Cars from Similar Training Images

Test Image

Likelihood of Car

\[ P_p(x_p = \text{car}) \]
Energy $E$ for Scene Parsing | Semantic Segmentation

$$E(X) = \sum_p D_p(x_p) + \lambda \sum_{p,q\in N} V_{pq}(x_p, x_q)$$

$P_p(x_p = c)$

$V_{pq}$ (sea, sand)
$V_{pq}$ (sky, rock)
$V_{pq}$ (sofa, sea)

© 2013 Adobe Systems Incorporated. All Rights Reserved.
Scene Parsing | Semantic Segmentation Results

Image

Human Annotation

Tighe CVPR13
Scene Parsing | Semantic Segmentation Results

Image

Human Annotation

Tighe CVPR13
Scene Parsing | Semantic Segmentation Results

Image

Human Annotation

Tighe CVPR13
Scene Parsing | Semantic Segmentation Results

Image

Human Annotation

Tighe CVPR13
Summary and Conclusion

- There are many variations of segmentation problems
  - How many classes? What are the classes?
  - What features are used?
  - Hard or Soft Segmentation?
  - Automatic or Interactive? What User Input?
  - How many images are segmented?

-Semantic Segmentation: still a lot of work to be done