Active and Interactive Image and Video Segmentation

Kristen Grauman
University of Texas at Austin

Work with Suyog Jain and Danna Gurari
Foreground Object Segmentation

**Task:** Generate pixel level masks for the foreground objects in an image or video
Why Foreground Object Segmentation?
Spectrum of automatic segmentation methods

Unsupervised methods


Fully supervised methods


Kristen Grauman, UT Austin
Symbiosis in Segmentation

Kristen Grauman, UT Austin
Symbiosis in Segmentation

High-level perception
Task awareness

Large-scale processing
Low-level details
Interactive Segmentation

Main idea in existing methods: Use “light” annotations to infer more precise boundaries


Kristen Grauman, UT Austin
Limiting assumptions in existing work

- One-size-fits-all annotation modalities
- Human always knows best
- Constant human in the loop to monitor video segmentation
Our goal

Active and interactive segmentation methods to predict exactly where and how human intervention is needed

This talk:

1. Given an image, what strength of annotation is needed?
2. Given a collection of images, which ones need human input?
3. Given a video, how to propagate minimal human input?

Kristen Grauman, UT Austin
Background: a typical MRF segmentation model

\[ E(L) = \sum_{p} A_p(y_p) + \sum_{p,q \in N} S_{p,q}(y_p, y_q) \]

\( y_p \in \{1, 0\} \) is the label of pixel \( p \)

User input leads to foreground likelihoods


Kristen Grauman, UT Austin
Problem

Fixing the input modality leads to a suboptimal trade-off between human and machine effort!

Kristen Grauman, UT Austin
I MUST BE WILLING TO TAKE ON EXTRA RESPONSIBILITIES?

FALSE. THAT'S NOT IN THE JOB DESCRIPTION.
Our Idea

Predict the annotation modality that is sufficiently strong for accurate segmentation

Expense vs. Precision

Bounding Box  Sloppy Contour  Tight Polygon

Low Cost  High Cost

Kristen Grauman, UT Austin

Jain & Grauman, ICCV 2013
Our Idea

Predict the annotation modality that is sufficiently strong for accurate segmentation

Expense vs. Precision

Bounding Box, Sloppy Contour, Tight Polygon

Low Cost, High Cost

Kristen Grauman, UT Austin

Jain & Grauman, ICCV 2013
Training Phase

• Given ground truth foreground, simulate the user input.

Ground Truth

Bounding Box

Sloppy Contour

Graph Cuts Segmentation

Compute object independent features which capture the foreground separability.
Training Phase: Learn Image Cues Indicative of Difficulty

- **Color Separability**
  
  ![Image](image1.png)  \[d = 0.6269\]  ![Image](image2.png)  \[d = 0.2764\]

- **Label Uncertainty**

- **Edge Complexity**

- **Boundary alignment**

Kristen Grauman, UT Austin  

Jain & Grauman, ICCV 2013
Training Phase:
Learn Image Cues Indicative of Difficulty

• Train model to predict difficulty for each input modality.

• Easiness = segmentation overlap score \( \frac{\text{Pred.} \cap \text{GT}}{\text{Pred.} \cup \text{GT}} \)

E.g. for bounding box:

Kristen Grauman, UT Austin

Jain & Grauman, ICCV 2013
Testing Phase: Will a Given Modality Succeed?

- Given novel image, salient object detector (Liu et al. 2009) to roughly localize probably foreground

Predict whether each modality would succeed:
1. Compute bounding boxes/sloppy contours from mask
2. Apply graph cut segmentation.
3. Extract features and predict the difficulty.

Kristen Grauman, UT Austin

Jain & Grauman, ICCV 2013
Datasets

- MSRC (591 images, 20 classes)
- CMU – Cornell iCoseg (643 images, 38 groups)
- Interactive Image Segmentation (151 unrelated images)
How well can we detect difficult images?

Our method learns generic cues, not dataset-specific features.

Kristen Grauman, UT Austin

Jain & Grauman, ICCV 2013
Qualitative Results – Success Cases

- **Bounding Box**
  - Sufficient
  - Examples: Panda, airplanes, kite, scissors, flower

- **Sloppy contour**
  - Sufficient
  - Examples: Bear, statue, nun, kendo, Taj Mahal, kangaroo

- **Tight Polygon**
  - Required
  - Examples: Elephants, motorcycle, car, leopard, cityscape
Qualitative Results – Failure Cases

Bounding Box sufficient

Sloppy contour sufficient

Tight Polygon required
Using difficulty predictions to intelligently gather annotations

1 Cascaded Selection

Kristen Grauman, UT Austin
For almost no loss in accuracy, our method leads to substantial savings in annotation effort.
Using difficulty predictions to intelligently gather annotations

1. Cascaded Selection

2. Budgeted Selection

Kristen Grauman, UT Austin
Using difficulty predictions to intelligently gather annotations

Given a cost budget, can we maximize the accuracy crowd will achieve in collaboration with algorithm?

101 Turkers contribute annotations
Kristen Grauman, UT Austin
Learning the failure behavior per segmentation algorithm

8 Algorithm-Drawn Segmentations

Ground Truth

Pinpoint which method new image should go to...or when to “pull the plug” and go to human annotator.

Kristen Grauman, UT Austin  Gurari et al. CVPR 2016
Our goal

Active and interactive segmentation methods to predict exactly where and how human intervention is needed

This talk:

1. Given an image, what strength of annotation is needed?

2. Given a collection of images, which ones need human input?

3. Given a video, how to propagate minimal human input?

Kristen Grauman, UT Austin
Symbiosis in Segmentation

**Traditional approach**: Propagate human input within the image.

Our goal: Active propagation

How to propagate human input segmentations across multiple images/frames?

Actively request human annotations for select images

Update segmentations for unlabeled images

Kristen Grauman, UT Austin
Key question 1: How to propagate?

Given some subset of labeled data, how to propagate to unlabeled data.
Weakly Supervised Scenario

Exploit repeated patterns by jointly segmenting out the foreground object
## Approach – Segmentation Propagation

Generate bottom-up object proposals for each image

<table>
<thead>
<tr>
<th>Images</th>
<th>Object proposals</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /> <img src="image2.png" alt="Image" /> <img src="image3.png" alt="Image" /></td>
<td><img src="object1.png" alt="Object Proposal" /> <img src="object2.png" alt="Object Proposal" /> <img src="object3.png" alt="Object Proposal" /></td>
</tr>
<tr>
<td><img src="image4.png" alt="Image" /> <img src="image5.png" alt="Image" /> <img src="image6.png" alt="Image" /></td>
<td><img src="object4.png" alt="Object Proposal" /> <img src="object5.png" alt="Object Proposal" /> <img src="object6.png" alt="Object Proposal" /></td>
</tr>
</tbody>
</table>

[Carreira 2012, Arbelaez 2014] [Jain & Grauman, CVPR 2016]
Approach – Segmentation Propagation

Goal: Select “good” proposals in each image

Images

Filtered object proposals

[Carreira 2012, Arbelaez 2014] [Jain & Grauman, CVPR 2016]
Approach – MRF Joint Segmentation

Unary Term: Average region saliency

\[ E(\mathcal{Y}) = \sum_{R_{ij}} - \log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij}) \]

[Jain & Grauman, CVPR 2016]
Approach – MRF Joint Segmentation

Pairwise connections between all region proposals

\[ E(\mathcal{Y}) = \sum_{R_{ij}} - \log \Phi(Y_{ij}) + \sum_{R_{ij}, R_{ij}' \in \mathcal{E}} \Psi(Y_{ij}, Y_{ij}') \]
Approach – MRF Joint Segmentation

Pairwise connections between all region proposals

\[ E(\mathcal{Y}) = \sum_{R_{ij}} - \log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij}) \]
Approach – MRF Joint Segmentation

Pairwise connections between all region proposals

\[ E(\mathcal{Y}) = \sum_{R_{ij}} - \log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij}) \]
Approach – MRF Joint Segmentation

Energy minimization using Graph-cuts

\[
E(\mathcal{Y}) = \sum_{R_{ij}} - \log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij})
\]
Approach – MRF Joint Segmentation

Energy minimization using Graph-cuts

Very efficient (1 min for 1400 images) compared to pixel based approach (225 hours) [Rubinstein 2012]

[Jain & Grauman, CVPR 2016]
Approach – MRF Joint Segmentation

Actively choose an image to be labeled by humans

\[ E(\mathcal{Y}) = \sum_{R_{ij}} - \log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij}) \]
Approach – MRF Joint Segmentation

Inject human-labeled regions in the joint graph

\[ E(\mathcal{Y}) = \sum_{R_{ij}} - \log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij}) \]
Weakly Supervised Segmentation

ImageNet dataset (~1M images, 3624 classes) [Deng 2009]

<table>
<thead>
<tr>
<th>Methods</th>
<th>ImageNet dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top obj. box [64]</td>
</tr>
<tr>
<td>BBox-CorLoc</td>
<td>37.42</td>
</tr>
</tbody>
</table>

We correctly localize 41,715 more images than next best approach.
Weakly Supervised Segmentation

MIT Object Discovery Dataset [Rubinstein 2012]

Consistently good performance that boosts state of the art in most cases

<table>
<thead>
<tr>
<th>Methods</th>
<th>MIT dataset (subset)</th>
<th></th>
<th>MIT dataset (full)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Airplane</td>
<td>Car</td>
<td>Horse</td>
<td>Airplane</td>
</tr>
<tr>
<td># Images</td>
<td>82</td>
<td>89</td>
<td>93</td>
<td>470</td>
</tr>
<tr>
<td>Joulin et al. [34]</td>
<td>15.36</td>
<td>37.15</td>
<td>30.16</td>
<td>n/a</td>
</tr>
<tr>
<td>Joulin et al. [35]</td>
<td>11.72</td>
<td>35.15</td>
<td>29.53</td>
<td>n/a</td>
</tr>
<tr>
<td>Kim et al. [37]</td>
<td>7.9</td>
<td>0.04</td>
<td>6.43</td>
<td>n/a</td>
</tr>
<tr>
<td>Rubinstein et al. [59]</td>
<td>55.81</td>
<td>64.42</td>
<td>51.65</td>
<td>55.62</td>
</tr>
<tr>
<td>Chen et al. [16]</td>
<td>54.62</td>
<td>69.2</td>
<td>44.46</td>
<td>60.87</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>58.65</strong></td>
<td><strong>66.47</strong></td>
<td><strong>53.57</strong></td>
<td><strong>62.27</strong></td>
</tr>
</tbody>
</table>
Key question 2: Which to annotate?

Given an annotation budget, which ones ought to be labeled by human annotators?

Kristen Grauman, UT Austin
Active Selection

Goal: Select a subset of images for human labeling
Active Selection

Diverse images
Active Selection

Uncertain images

Predict quality of current foreground estimate
Active Segmentation Propagation

ImageNet Dataset

MIT Object Discovery Dataset

Random  PageRank  Ours without uncertainty  Ours

[Rubinstein 2012]
Our goal

**Active and interactive segmentation methods to predict exactly where and how human intervention is needed**

This talk:

1. **Given an image**, what strength of annotation is needed?

2. **Given a collection of images**, which ones need human input?

3. **Given a video**, how to propagate minimal human input?

Kristen Grauman, UT Austin
Propagation in Video: Problem

Existing methods [Tsai 2010, Fathi 2011, Vijayanarasimhan 2012] can only enforce local consistency in space and time (using pairwise connections).

Robust foreground propagation requires capturing long range dependencies as object evolves in shape over time.

Kristen Grauman, UT Austin
Propagation in video

Supervoxels: bottom-up space-time regions
[Grundmann 2010, Xu 2012]

**Our idea:** Higher order potentials over supervoxels to enforce long term temporal consistency

S. Jain, K. Grauman, Supervoxel-Consistent Foreground Propagation in Video, ECCV 2014
Kristen Grauman, UT Austin
Propagation in video

\[ E(\mathcal{Y}) = \sum_{(t,i) \in \mathcal{X}} \Phi^i_t(y^i_t) + \sum_{[(t,i),(t',j)] \in \mathcal{E}, t' \in \{t,t+1\}} \Phi^i,j_{t,t'}(y^i_t, y^j_{t'}) + \sum_{v \in S} \Phi_v(y_v) \]

Assign soft preferences for label consistency within supervoxels
Robust P^n model [Kohli 2008]
Results

**Video**

**PF-MRF**
[Vijayanarasimhan 2012]

**Supervoxels**

**Ours**
[Jain & Grauman, ECCV 2014]
Click Carving for video segmentation

- Interactively segment the frame to be propagated: boundary clicks fetch relevant object proposals

Click Carving: Segmenting Objects in Video with Point Clicks

[Jain & Grauman, HCOMP 2016]
Click Carving for video segmentation

• Results achieved with average of 2 user clicks

[Jain & Grauman, HCOMP 2016]
Active human-machine collaboration for foreground object segmentation in images and video

- Active selection of sufficiently strong annotation modality to initialize interactive image segmentation
- Active segmentation propagation for large weakly supervised image collections
- Click carving and high order supervoxel potentials for segmentation propagation in video

Kristen Grauman, UT Austin
References


