Predicting Sufficient Annotation Strength for Interactive Foreground Segmentation
Suyog Jain and Kristen Grauman
University of Texas at Austin

Interactive Image Segmentation
Human provides high level guidance to the segmentation algorithm.
Mobile Search Data Collection Graphics

Problem: Fixing the input modality for interactive segmentation methods is not optimal
Bounding Box Sloppy Contour

Our Goal
Predict the annotation modality that is sufficiently strong for accurate segmentation of a given image.

Low Cost High Cost

Applications:
Quick selection for a single image Group selection with fixed budget

$\text{Cost} = \sum_{p} A_{p}(y_{p}) + \sum_{p \in N^{S}} S_{p,q}(y_{p}, y_{q})$

(Data term) (Smoothness term)
y_{p} \in \{0,1\}$ is the label of pixel p$
L : \text{Labeling over entire image}$

Boykov 2001, Rother 2004

Learning to predict segmentation difficulty per modality
Training: Given a set of images with the foreground masks, we simulate the user input
Segmentation with simulated user input

Image Bounding Box Sloppy Contour

Fit a tight rectangle Dilate ground truth

Object independent features
Color distances Graph Cuts Uncertainty Edge histogram Boundary alignment

Use the overlap score between the resulting segmentation and ground truth to mark an image as "easy" or "hard" and train a linear SVM classifier (for each modality).

Easy Hard
Testing: Use saliency detector to get a coarse estimate of foreground at test time.

Compute the proposed features and use trained classifiers to predict difficulty.

Audit Box? Sloppy Contour? Tight Polygon?

Cascade selection

Goal: Given a batch of "n" images with a fixed time budget "B", we find the optimal annotation tool for each image.

Objective: $x^{*} = \arg \max_{x} \sum_{k=1}^{n} p_{k} x_{k}^{\text{mod}} + p_{k} x_{k}^{\text{basic}} + p_{k} x_{k}^{\text{phys}}$

s.t. $c^{T} x \leq B_{k}$

(Selection should not exceed budget)

Constraints

$x_{k}^{\text{mod}} + x_{k}^{\text{basic}} + x_{k}^{\text{phys}} = 1, \forall k \in 1, \ldots, n$ (Uniqueness Constraint)

$x_{k}^{\text{mod}}, x_{k}^{\text{basic}}, x_{k}^{\text{phys}} \in \{0,1\}, \forall k \in 1, \ldots, n$

where, $x$: Modality annotation vector $p$: Success probability $c$: Cost vector

Efficiently solved using branch and bound method.

Results
Predicting segmentation difficulty per modality
Bounding box sufficient Sloppy contour sufficient Tight polygon required

Datasets:
- MSRC (591 images)
- iCoseg (643 images)
- IIS (151 images)
Leave one dataset out cross-validation.

Our method learns generic cues to predict difficulty, not some dataset specific properties.

Baselines:
- Otsu adaptive thresholding
- SVM with global image features
- Our method with Ground Truth input (upper bound)

Cascade selection – application to object recognition
Task: Given a set of images with a common object, train a classifier to separate object vs. non object regions.

How to get data labeled?
All tight: Ask the human annotator to provide pixel level masks (status quo).
Ours: Use our cascade selection method to decide the best annotation for each image.

Annotation choices with budget constraints
Goal: Given a batch of "n" images with a fixed time budget "B", we find the optimal annotation tool for each image.

Objective: $x^{*} = \arg \max_{x} \sum_{k=1}^{n} p_{k} x_{k}^{\text{mod}} + p_{k} x_{k}^{\text{basic}} + p_{k} x_{k}^{\text{phys}}$

s.t. $c^{T} x \leq B_{k}$

(Selection should not exceed budget)

Constraints

$x_{k}^{\text{mod}} + x_{k}^{\text{basic}} + x_{k}^{\text{phys}} = 1, \forall k \in 1, \ldots, n$ (Uniqueness Constraint)

$x_{k}^{\text{mod}}, x_{k}^{\text{basic}}, x_{k}^{\text{phys}} \in \{0,1\}, \forall k \in 1, \ldots, n$

where, $x$: Modality annotation vector $p$: Success probability $c$: Cost vector

Efficiently solved using branch and bound method.

Annotation choices with budget constraints (MTurk user study)

- 101 MTurkers (5 per image).
- Use the median time for each image for experiments.
- Budget ranges from "all bounding boxes" to "all tight polygons"

For the same amount of annotation time, our method leads to much higher average overlap scores.

Conclusion
A method to predict the kind of human annotation required to segment a given image.
User study shows that explicit reasoning about segmentation difficulty is useful.