1. Problem: Instance Segmentation

Given a depth image consisting of multiple non-rigid instances of an approximately convex object, our task is to predict the segmentation masks for the instances. We assume to have access to only a few annotated training examples.

Why not predict all instance segmentations? We consider a robotic bin-picking setting where the robot picks one instance at a time. Thus, segmenting some pickable instances is sufficient, and once those are picked up, we could iteratively apply the method to the remaining instances.

2. Contributions

- We present a simple approach – discriminative 3D shape modeling using surface geodesics – for instance segmentation in depth images.
- Our method needs only a few annotated examples to train our model, is very fast to train and predict (10 min to train and 0.1s to predict using a CPU).
- Our method shows large-margin improvements in instance segmentation performance on our challenging Food-Items dataset.

3. Prior Works

- Supervised Methods: Mask-RCNN and related methods
  - Needs many annotated training examples, could be expensive or challenging to gather in changing real-world conditions.
- Unsupervised Methods: InSegAN, Slot Attention, IODINE, etc.
  - Needs a large-sized (e.g., thousands of images) of unannotated data
  - May be time consuming
- Our method needs only few-shot annotated examples to train our model.

4. Discriminative 3D Shapes Using Depth Geodesics

Key idea: Take any two points on the depth surface and draw a surface geodesic/curve between the points. Can we learn the similarities in the geodesic as it traverses on the instance's surface to predict if its end-points belong to the same instance or multiple instances? We assume the 3D shape is approximately convex.

4a. Training Pipeline

We use the trained neural network for classifying the geodesics.

4b. Inference Pipeline

5. Experiments: Food-Items Dataset

- Four object categories
- Captured using Ensenso depth camera
- 20 annotated images per category
- 2–30 instances per image
- Used 3 images for training per category
- Used a simple neural network with 5 MLP layers.
- Used mIoU for evaluation using ground truth annotations.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cat:007</th>
<th>Cat:111</th>
<th>Chicken</th>
<th>Fried Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>KMeans</td>
<td>0.510</td>
<td>0.481</td>
<td>0.480</td>
<td>0.435</td>
</tr>
<tr>
<td>GMM (std)</td>
<td>0.510</td>
<td>0.490</td>
<td>0.494</td>
<td>0.442</td>
</tr>
<tr>
<td>Spectral</td>
<td>0.475</td>
<td>0.464</td>
<td>0.477</td>
<td>0.572</td>
</tr>
<tr>
<td>Waterdip [14]</td>
<td>0.687</td>
<td>0.539</td>
<td>0.585</td>
<td>0.862</td>
</tr>
<tr>
<td>LCIP [26]</td>
<td>0.495</td>
<td>0.457</td>
<td>0.440</td>
<td>0.501</td>
</tr>
<tr>
<td>SIFT [12]</td>
<td>0.420</td>
<td>0.357</td>
<td>0.370</td>
<td>0.429</td>
</tr>
<tr>
<td>MSIO [16]</td>
<td>0.761</td>
<td>0.532</td>
<td>0.790</td>
<td>0.844</td>
</tr>
<tr>
<td>Mask-RCNN</td>
<td>0.659</td>
<td>0.712</td>
<td>0.591</td>
<td>0.262</td>
</tr>
</tbody>
</table>

Results using Mask-RCNN (see false positives and over-segmentations?)

6. Discriminative 3D Shape Modeling for Few-Shot Instance Segmentation

Anoop Cherian Siddharth Jain Tim K. Marks Alan Sullivan
Mitsubishi Electric Research Labs (MERL), Cambridge, MA
{cherian, sjain, tmarks, sullivan}@merl.com