What is this pixel?
Segmentation and Context

What is this pixel?
Segmentation and Context

What is this pixel?
Now what is this pixel?
Now what is this pixel?
Segmentation and Context
Goal of Scene Decomposition

- Decompose the scene into **regions** with
  - semantic region labels (e.g., road, sky, building, etc.)
  - coherent geometric placement (orientation and location with respect to the horizon)
Region-based Model

Variables
- $\alpha_p$: pixel appearance
- $R_p$: pixel-to-region correspondence
- $A_r$: region appearance
- $S_r$: region semantic class
- $G_r$: region geometry
- $v^{hz}$: location of horizon

Energy Function
$$E(R, A, S, G, v^{hz}, K | I, \theta)$$
Energy Function

\[ E(R, A, S, G, \nu_{hz}, K \mid I, \theta) = \psi_{\text{horizon}}(\nu_{hz}) + \psi_{\text{region}}(S_r, G_r, A_r, \nu_{hz}) + \psi_{\text{boundary}}(A_r, A_s) + \psi_{\text{pair}}(S_r, S_s, G_r, G_s) \]

- **Horizon Term**: e.g., vanishing lines
- **Region Term**: e.g., consistent appearance and location
- **Boundary Term**: e.g., difference in color/texture between regions
- **Pairwise Term**: e.g., foreground on road
Energy Function

\[ E(R, A, S, G, \psi^{hz}, K \mid I, \theta) \]

\[ = \]

\[ \psi^{\text{horizon}}(\psi^{hz}) \quad \psi^{\text{region}}(S_r, G_r, A_r, \psi^{hz}) \quad \psi^{\text{boundary}}(A_r, A_s) \quad \psi^{\text{pair}}(S_r, S_s, G_r, G_s) \]

**Exact inference is intractable**
Inference

image

segment database ($\Omega$)

initialize

scene decomposition

proposal move ($R_p$)
(Segment) Proposal Moves

- Initial decomposition
- Proposal move
- Final decomposition

Segment database ($\Omega$)
Inference

image

segment database ($\Omega$)

initialize

scene decomposition

proposal move ($R_p$)

global inference ($S_r, G_r, v^{hz}$)

accept if lower

$$E(R, A, S, G, v^{hz}, K | I, \theta)$$

evaluate energy function
Inference Animation

Decomposing a Scene into Geometric and Semantically Consistent Regions

Stephen Gould
Daphne Koller

International Conference on Computer Vision 2009
Parameter Learning

- Positive examples: all coherent regions and segments
- Negative examples: exponentially many
  - Most of them are ridiculously easy
- Closed-loop learning
  - Learn simple region and context models
  - Run inference (on training set) sampling errors
  - Re-train with augmented training set

\[ E(R, A, S, G, v^c, K | I, 0) \]
Results: 21-class MSRC

- Validate against state-of-the-art approaches
- Region/pixel class only
- Ground truth labels are approximate
- **No geometry** information

<table>
<thead>
<tr>
<th>21 CLASS</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shotton et al.</td>
<td>72.2</td>
</tr>
<tr>
<td>Gould et al.</td>
<td><strong>76.5</strong></td>
</tr>
<tr>
<td>Pixel CRF</td>
<td>75.3</td>
</tr>
<tr>
<td>Region-based</td>
<td>76.4</td>
</tr>
</tbody>
</table>

![Images of cows and sheep with annotations](image-url)
High Quality Dataset

- MSRC dataset is limited
  - poorly labeled boundaries
  - many missing pixels (void)
  - no geometry information

- Collected images from MSRC, Hoiem et al., Pascal VOC

- 715 outdoor scenes with high-quality labels
  - region boundaries
  - region class and geometry
  - horizon

- Used Amazon’s Mechanical Turk for labeling

- Available for download from: http://www.stanford.edu/~sgould
$0.10 per task (regions, classes, surface types)
5-10 minutes per task
24-48 hour turn-around time (for 715 images)
Less than 10% of tasks needed rework

**Total cost for labels:** under $250 (includes $40 textbook on Adobe Flash)

**Saving me from having to label image:** priceless.
AMT: Label Quality

You don’t always get what you want

Typical quality (hand labeled)

Comparison with MSRC labels
Quantitative Results

### CLASS

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel CRF</td>
<td>74.3</td>
<td>0.80</td>
</tr>
<tr>
<td>Region-based</td>
<td><strong>76.4</strong></td>
<td>1.22</td>
</tr>
</tbody>
</table>

**Region Classes:** sky, tree, road, grass, water, building, mountain, fg. object

### GEOMETRY

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel CRF</td>
<td>89.1</td>
<td>0.73</td>
</tr>
<tr>
<td>Region-based</td>
<td><strong>91.0</strong></td>
<td>0.56</td>
</tr>
</tbody>
</table>

**Region Geometry:** sky, vertical, horizontal (support)

**Horizon error:** 6.9% (17 pixels)
Example Results
Application: 3d Reconstruction

- Estimate camera tilt from location of horizon
- Predict region 3d position using ray projected through camera plane
Example 3d Reconstructions

Related work: [Saxena et al., PAMI 08], [Hoiem et al., IJCV 07], [Russell and Torralba, CVPR 09]
NI PS 2009 Sneak Peak
Hierarchical Scene Model

Variables
- $\alpha_p$: pixel appearance
- $R_p$: pixel-to-region correspondence
- $A_r$: region appearance
- $S_r$: region semantic class
- $G_r$: region geometry
- $O_r$: region-to-object correspondence
- $C_o$: object class
- $v^{hz}$: location of horizon

Energy function
Energy Function

$$E(R, A, S, G, O, C, v^{hz}, K \mid I, \theta)$$

- **Horizon Term**
  - e.g., vanishing lines
  - $\psi_{\text{horizon}}(v^{hz})$

- **Region Term**
  - e.g., consistent appearance and location
  - $\psi_{\text{region}}(S_r, G_r, v^{hz})$

- **Boundary Term**
  - e.g., difference in color/texture between regions
  - $\psi_{\text{boundary}}(A_r, A_s)$

- **Object Term**
  - e.g. wheel-like appearance in bottom corner
  - $\psi_{\text{object}}(C_o, v^{hz})$

- **Context Term**
  - e.g., cars on road
  - $\psi_{\text{context}}(C_o, S_r)$
Top-down Proposal Moves

- Image
- Initialize
- Scene decomposition
- Proposal move \((R_p, O_r)\)
- Global inference \((S_r, G_r, C_o, v^{hz})\)

\[ E(R, A, S, G, O, C, v^{hz}, K | I, \theta) \]

Evaluate energy function

Gould, et al.
Object Detection Results

Sliding-window detector’s top two detections per image

- Person (88.1%)
- Person (98.1%)
- Car (88.3%)
- Car (88.9%)
- Cow (98.4%)

Our joint region-based segmentation and object detection
Our model decomposes a scene into geometric and semantically consistent regions using a unified energy function over pixels and regions.

By classifying large regions rather than individual pixels we can compute more robust features and reduce inference complexity.

Multiple over-segmentations allow us to refine region boundaries and make large moves in energy space.

Context can be easily captured using a pairwise term between adjacent regions.

Our model provides a base for integrating many other vision tasks (e.g., 3D reconstruction and object detection).
Thank You ありがとうとうございます。