Resolving Stereo Ambiguities using Object Knowledge

Andreas Geiger
Fatma Güney

MPI for Intelligent Systems
Tübingen
Stereo Matching

Task:

- Given a rectified stereo image pair
- Estimate the disparity (≡displacement) at each pixel
Stereo Matching

Task:
- Given a rectified stereo image pair
- Estimate the disparity (≡displacement) at each pixel
Stereo Matching

Energy Minimization Problem:

\[ E = \sum_i E_{data}(d_i) + \sum_{i \sim j} E_{smooth}(d_i, d_j) \]

- \( d_i \): disparity of pixel \( i \) or plane parameters of superpixel \( i \)
- \( i \sim j \): neighbors
Stereo Matching

Question:
- Is the stereo problem solved?
- How can we address textureless or specular surfaces?

Key idea:
- Introduce top-down object knowledge into the process!
Question:
- Is the stereo problem solved?
- How can we address textureless or specular surfaces?

Key idea:
- Introduce top-down object knowledge into the process!
Stereo Matching

Question:
- Is the stereo problem solved?
- How can we address textureless or specular surfaces?

Key idea:
- Introduce top-down object knowledge into the process!
Stereo Matching

Question:
- Is the stereo problem solved?
- How can we address textureless or specular surfaces?

Key idea:
- Introduce top-down object knowledge into the process!
High-order Models in Computer Vision

Image Denoising

Image Segmentation

[Roth & Black, 2009]

[Kohli, Ladicky & Torr, 2009]
Overview

Input Images → Superpixels & Semantics

Initial Disparity Map → 3D CAD

Final Result → Inference

Displets → Inverse Graphics
Overview
Initial Disparity Map

- CNN-based stereo matching cost [Zbontar et al., 2015]
- Semi-global matching [Hirschmüller, 2008]
- Left-right consistency check
Superpixels and Semantic Segmentation

- StereoSLIC superpixels [Yamaguchi et al., 2013]
- Associative Hierarchical Random Fields for vehicle vs. background segmentation [Ladicky et al., 2013]
Rapid Inverse Graphics

- **Goal:** Generate plausible 3D geometries for a given semantic class
- MCMC simulation based on 3D CAD models (∼8200 fps)
Semi-Convex Hull
Displets
Inference

Goal: Minimize Gibbs energy ...

\[ E = \sum_{i \in S} \varphi_i^S (n_i) + \sum_{i \sim j} \psi_{ij}^S (n_i, n_j) + \sum_{k \in D} \varphi_k^D (d_k) + \sum_{k \in D} \sum_{i \in S_k} \psi_{ki}^D (d_k, n_i) \]

... with respect to superpixel normals \( n_i \in \mathbb{R}^3 \) and displets \( d_k \in \{0, 1\} \)

Notation:
- \( S \): set of all superpixels
- \( D \): set of all displets
- \( S_k \): set of superpixels covered by displet \( k \)

Inference:
- Max-Product Particle Belief Propagation (MP-PBP)
Inference

Goal: Minimize Gibbs energy ...

\[ E = \sum_{i \in S} \phi^S_i(n_i) + \sum_{i \sim j} \psi^S_{ij}(n_i, n_j) + \sum_{k \in D} \phi^D_k(d_k) + \sum_{k \in D} \sum_{i \in S_k} \psi^D_{ki}(d_k, n_i) \]

... with respect to superpixel normals \( n_i \in \mathbb{R}^3 \) and displets \( d_k \in \{0, 1\} \)

Data Term:
- Penalizes deviation from initial disparity map (truncated \( \ell_1 \))
- Measures photo-consistency between left and right image
Inference

Goal: Minimize Gibbs energy ...

$$E = \sum_{i \in S} \varphi_i^S(n_i) + \sum_{i \sim j} \psi_{ij}^S(n_i, n_j) + \sum_{k \in D} \varphi_k^D(d_k) + \sum_{k \in D} \sum_{i \in S_k} \psi_{ki}^D(d_k, n_i)$$

... with respect to superpixel normals $n_i \in \mathbb{R}^3$ and displets $d_k \in \{0, 1\}$

Smoothness Term:
- Penalizes depth discontinuities at superpixel boundaries
- Encourages similar orientations of neighboring superpixel planes
Inference

Goal: Minimize Gibbs energy ...

\[ E = \sum_{i \in S} \varphi^S_i(n_i) + \sum_{i \sim j} \psi^S_{ij}(n_i, n_j) + \sum_{k \in D} \varphi^D_k(d_k) + \sum_{k \in D} \sum_{i \in S_k} \psi^D_{ki}(d_k, n_i) \]

... with respect to superpixel normals \( n_i \in \mathbb{R}^3 \) and displets \( d_k \in \{0, 1\} \)

Displet Unary Term:

- Encourages image regions with semantic class label \( c_k \) to be explained by a displet of the corresponding class:

\[ \varphi^D_k(d_k) = -\theta_3 [d_k = 1] \cdot (|S = c_k| \cap M_k| + \kappa_k) \]

- \( M_k \): foreground mask of displet \( k \)  \( S \): semantic segmentation
Inference

**Goal:** Minimize Gibbs energy ...

\[
E = \sum_{i \in S} \varphi^S_i(n_i) + \sum_{i \sim j} \psi^S_{ij}(n_i, n_j) + \sum_{k \in D} \varphi^D_k(d_k) + \sum_{k \in D} \sum_{i \in S_k} \psi^D_{ki}(d_k, n_i)
\]

... with respect to superpixel normals \( n_i \in \mathbb{R}^3 \) and displets \( d_k \in \{0, 1\} \)

**Displet Consistency Term:**
- Encourages superpixels to agree with the geometry of the displets:

\[
\psi^D_{ki}(d_k, n_i) = \lambda_{ki} [d_k = 1] \cdot (1 - \delta(n_i, \hat{n}_k, z_i))
\]
- \( \lambda_{ki} \): weight depending on distance to boundary:
## Results on KITTI Validation Set (Ablation)

<table>
<thead>
<tr>
<th>Model Configuration</th>
<th>Reflective Regions</th>
<th>All Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input $\hat{\Omega}$ (Interpolated)</td>
<td>19.84 %</td>
<td>3.35 %</td>
</tr>
<tr>
<td>Unary Only</td>
<td>17.72 %</td>
<td>3.31 %</td>
</tr>
<tr>
<td>Unary + Pair (Boundary)</td>
<td>15.96 %</td>
<td>3.21 %</td>
</tr>
<tr>
<td>Unary + Pair (Normal)</td>
<td>17.06 %</td>
<td>3.28 %</td>
</tr>
<tr>
<td>Unary + Pair</td>
<td>14.78 %</td>
<td>3.12 %</td>
</tr>
<tr>
<td>Unary + Pair + Occ</td>
<td>15.32 %</td>
<td>3.04 %</td>
</tr>
<tr>
<td>Unary + Pair + Disp</td>
<td>7.08 %</td>
<td>2.87 %</td>
</tr>
<tr>
<td>Unary + Pair + Occ + Disp</td>
<td>7.16 %</td>
<td>2.78 %</td>
</tr>
</tbody>
</table>

- **Metric**: Percentage of erroneous pixels (> 3px disparity error)
### Results on KITTI Test Set (Reflective Regions)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Method</th>
<th>Out-Noc</th>
<th>Out-All</th>
<th>Avg-Noc</th>
<th>Avg-All</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Our Method</td>
<td>8.40 %</td>
<td>9.89 %</td>
<td>1.9 px</td>
<td>2.3 px</td>
</tr>
<tr>
<td>2</td>
<td>VC-SF *</td>
<td>11.58 %</td>
<td>12.29 %</td>
<td>2.7 px</td>
<td>2.8 px</td>
</tr>
<tr>
<td>3</td>
<td>PCB-P-SS</td>
<td>14.26 %</td>
<td>18.33 %</td>
<td>2.4 px</td>
<td>3.9 px</td>
</tr>
<tr>
<td>4</td>
<td>SPS-StFl *</td>
<td>14.74 %</td>
<td>18.00 %</td>
<td>2.9 px</td>
<td>3.6 px</td>
</tr>
<tr>
<td>5</td>
<td>CoP</td>
<td>15.30 %</td>
<td>19.15 %</td>
<td>2.7 px</td>
<td>4.1 px</td>
</tr>
<tr>
<td>6</td>
<td>SPS-St</td>
<td>16.05 %</td>
<td>19.34 %</td>
<td>3.1 px</td>
<td>3.6 px</td>
</tr>
<tr>
<td>7</td>
<td>DDS-SS</td>
<td>16.23 %</td>
<td>19.39 %</td>
<td>2.5 px</td>
<td>3.0 px</td>
</tr>
<tr>
<td>8</td>
<td>PCB-P</td>
<td>16.28 %</td>
<td>20.22 %</td>
<td>2.8 px</td>
<td>4.4 px</td>
</tr>
<tr>
<td>9</td>
<td>PR-Sf+E *</td>
<td>17.85 %</td>
<td>20.82 %</td>
<td>3.3 px</td>
<td>4.0 px</td>
</tr>
<tr>
<td>10</td>
<td>StereoSLIC</td>
<td>18.22 %</td>
<td>21.60 %</td>
<td>2.8 px</td>
<td>3.6 px</td>
</tr>
<tr>
<td>11</td>
<td>MC-CNN</td>
<td>18.45 %</td>
<td>21.96 %</td>
<td>3.5 px</td>
<td>4.3 px</td>
</tr>
<tr>
<td>12</td>
<td>PR-Sceneflow *</td>
<td>19.22 %</td>
<td>22.07 %</td>
<td>3.3 px</td>
<td>4.0 px</td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>62</td>
<td>ALE-Stereo</td>
<td>83.80 %</td>
<td>84.37 %</td>
<td>24.6 px</td>
<td>25.4 px</td>
</tr>
</tbody>
</table>
## Results on KITTI Test Set (All Regions)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Method</th>
<th>Out-Noc</th>
<th>Out-All</th>
<th>Avg-Noc</th>
<th>Avg-All</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Our Method</td>
<td>2.47 %</td>
<td>3.27 %</td>
<td>0.7 px</td>
<td>0.9 px</td>
</tr>
<tr>
<td>2</td>
<td>MC-CNN</td>
<td>2.61 %</td>
<td>3.84 %</td>
<td>0.8 px</td>
<td>1.0 px</td>
</tr>
<tr>
<td>3</td>
<td>SPS-StFl *</td>
<td>2.83 %</td>
<td>3.64 %</td>
<td>0.8 px</td>
<td>0.9 px</td>
</tr>
<tr>
<td>4</td>
<td>VC-SF *</td>
<td>3.05 %</td>
<td>3.31 %</td>
<td>0.8 px</td>
<td>0.8 px</td>
</tr>
<tr>
<td>5</td>
<td>SPS-St</td>
<td>3.39 %</td>
<td>4.41 %</td>
<td>0.9 px</td>
<td>1.0 px</td>
</tr>
<tr>
<td>6</td>
<td>PCBP-SS</td>
<td>3.40 %</td>
<td>4.72 %</td>
<td>0.8 px</td>
<td>1.0 px</td>
</tr>
<tr>
<td>7</td>
<td>CoP</td>
<td>3.78 %</td>
<td>4.63 %</td>
<td>0.9 px</td>
<td>1.1 px</td>
</tr>
<tr>
<td>8</td>
<td>DDS-SS</td>
<td>3.83 %</td>
<td>4.59 %</td>
<td>0.9 px</td>
<td>1.0 px</td>
</tr>
<tr>
<td>9</td>
<td>StereoSLIC</td>
<td>3.92 %</td>
<td>5.11 %</td>
<td>0.9 px</td>
<td>1.0 px</td>
</tr>
<tr>
<td>10</td>
<td>PR-Sf+E *</td>
<td>4.02 %</td>
<td>4.87 %</td>
<td>0.9 px</td>
<td>1.0 px</td>
</tr>
<tr>
<td>11</td>
<td>PCBP</td>
<td>4.04 %</td>
<td>5.37 %</td>
<td>0.9 px</td>
<td>1.1 px</td>
</tr>
<tr>
<td>12</td>
<td>PR-SceneFlow *</td>
<td>4.36 %</td>
<td>5.22 %</td>
<td>0.9 px</td>
<td>1.1 px</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>62</td>
<td>ALE-Stereo</td>
<td>50.48 %</td>
<td>51.19 %</td>
<td>13.0 px</td>
<td>13.5 px</td>
</tr>
</tbody>
</table>
Qualitative Results: Success Cases
Qualitative Results: Success Cases
Qualitative Results: Success Cases
Qualitative Results: Success Cases
Qualitative Results: Failure Cases
Qualitative Results: Failure Cases
Qualitative Results: Video

MC-CNN (Zbontar et al.)

Our Results
Thank you!