Depth and Surface Normal Estimation from a Single Image

Mian Wei
University of Toronto
What is the problem?
Given one image
Estimate the following:
Eigen, D. and Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. ICCV 2015
Eigen, D. and Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. ICCV 2015
Why is this hard?
Multiple ambiguities
Scale ambiguity
Bas-relief ambiguity

Let’s play a game
Spot the Difference
All the same
Family of transformation
Generalized Bas-Relief
Change shape and illumination
Yield same image
Existing works
Multi-view Stereo

Photometric Stereo

Woodham, R.J. (1980), Photometric method for determining surface orientation from multiple images, Optical Engineering 19 (1) 139-144.
Collimated Light Sources
Light rays parallel
Shape from Focus

Light Fall-off Stereo

Specialized Hardware
Laser Scanner
Active Illumination
Time of Flight
Estimating Depth

Train 2 networks
Global coarse-scale network
Local fine-scale network
Global coarse-scale network
Learns a coarse depth map
Used as input to local network
Intuition:
Coarse info learnt already
Focus on learning finer info
Scale ambiguity
Scale invariant error function
$$D(y, y^*) = \frac{1}{2n} \sum_{i=1}^{n} (\log y_i - \log y_i^* + \alpha(y_i, y_i^*))^2$$

$$\alpha(y_i, y_i^*) = \frac{1}{n} \sum_{i=1}^{n} (\log y_i^* - \log y_i)$$
\[ D(ay, ay^*) = \frac{1}{2n} \sum_{i=1}^{n} (\log ay_i - \log ay_i^* + \alpha(ay_i, ay_i^*))^2 \]

\[ D(ay, ay^*) = \frac{1}{2n} \sum_{i=1}^{n} (\log a - \log a + \log y_i - \log y_i^* + \alpha(ay_i, ay_i^*))^2 \]

\[ D(ay, ay^*) = \frac{1}{2n} \sum_{i=1}^{n} (\log y_i - \log y_i^* + \log a - \log a + \alpha(y_i, y_i^*))^2 \]

\[ D(ay, ay^*) = D(y, y^*) \]
Loss Function
Scale invariant
$$L(y, y^*) = \frac{1}{n} \sum_{i=1}^{n} d_i^2 - \frac{\lambda}{n^2} \left( \sum_{i=1}^{n} d_i \right)^2$$

$$d_i = \log y_i - \log y_i^*$$
2 Datasets
NYUDepthV2

Indoor Rooms
Outdoor images taken on a car
How do you get ground truth?
NYUDepthV2
Kinect
KITT1
Time of Flight
Times how long light travels
From light source to camera
Results
<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Make3D</th>
<th>Ladicky &amp; al</th>
<th>Karsch &amp; al</th>
<th>Coarse</th>
<th>Coarse + Fine</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>threshold $\delta &lt; 1.25$</td>
<td>0.418</td>
<td>0.447</td>
<td>0.542</td>
<td>–</td>
<td>0.618</td>
<td>0.611</td>
<td></td>
</tr>
<tr>
<td>threshold $\delta &lt; 1.25^2$</td>
<td>0.711</td>
<td>0.745</td>
<td>0.829</td>
<td>–</td>
<td>0.891</td>
<td>0.887</td>
<td></td>
</tr>
<tr>
<td>threshold $\delta &lt; 1.25^3$</td>
<td>0.874</td>
<td>0.897</td>
<td>0.940</td>
<td>–</td>
<td>0.969</td>
<td>0.971</td>
<td></td>
</tr>
<tr>
<td>abs relative difference</td>
<td>0.408</td>
<td>0.349</td>
<td>–</td>
<td>0.350</td>
<td>0.228</td>
<td>0.215</td>
<td></td>
</tr>
<tr>
<td>sqr relative difference</td>
<td>0.581</td>
<td>0.492</td>
<td>–</td>
<td>–</td>
<td>0.223</td>
<td>0.212</td>
<td>lower is better</td>
</tr>
<tr>
<td>RMSE (linear)</td>
<td>1.244</td>
<td>1.214</td>
<td>–</td>
<td>1.2</td>
<td>0.871</td>
<td>0.907</td>
<td></td>
</tr>
<tr>
<td>RMSE (log)</td>
<td>0.430</td>
<td>0.409</td>
<td>–</td>
<td>–</td>
<td>0.283</td>
<td>0.285</td>
<td></td>
</tr>
<tr>
<td>RMSE (log, scale inv.)</td>
<td>0.304</td>
<td>0.325</td>
<td>–</td>
<td>–</td>
<td>0.221</td>
<td>0.219</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparison on the NYUDepth dataset
Estimating Surface Normals

Similar to Eigen
Trains 3 networks
Global coarse-scale network
Trains for room layout as well
Local fine-scale network
195x260x3 Image

Conv Layers

4 Layers

(a)

(b)
Trains for edge labels as well
Convex, concave, occlusion, N/A
Difference: Global and Local trained separately
Fusion Network
Combines both networks
How to represent normals
Normals lie in continuous space
Regression as Classification
Surface normal triangular coding
Codebook with k-means
Delaunay Triangulation cover
Triangles as classes
Represent Surface Normals
Weighted sum of triangle corners
Loss Function
\[ L(I, Y) = - \sum_{i=1}^{M \times M} \sum_{k=1}^{K} (1(y_i = k) \log F_{i,k}(I)) \]
Table 2: Ablative Analysis

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>RMSE 11.25°</th>
<th>RMSE 22.5°</th>
<th>RMSE 30°</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>25.0</td>
<td>13.8</td>
<td>35.9</td>
<td>44.2</td>
<td>63.2</td>
</tr>
<tr>
<td>Full (Soft)</td>
<td>24.2</td>
<td>17.3</td>
<td>32.2</td>
<td>36.8</td>
<td>58.5</td>
</tr>
<tr>
<td>Fusion (+VP)</td>
<td>25.3</td>
<td>14.4</td>
<td>35.9</td>
<td>42.7</td>
<td>62.5</td>
</tr>
<tr>
<td>Fusion (+Edge)</td>
<td>25.8</td>
<td>15.3</td>
<td>36.0</td>
<td>40.0</td>
<td>61.6</td>
</tr>
<tr>
<td>Fusion (+Layout)</td>
<td>25.8</td>
<td>14.9</td>
<td>36.3</td>
<td>41.1</td>
<td>61.9</td>
</tr>
<tr>
<td>Fusion</td>
<td>26.0</td>
<td>15.5</td>
<td>36.2</td>
<td>39.5</td>
<td>61.3</td>
</tr>
<tr>
<td>Bottom-up</td>
<td>32.2</td>
<td>23.5</td>
<td>42.0</td>
<td>27.2</td>
<td>48.5</td>
</tr>
<tr>
<td>Top-down</td>
<td>29.0</td>
<td>19.8</td>
<td>38.3</td>
<td>32.7</td>
<td>53.8</td>
</tr>
<tr>
<td>Eigen et al. (Fusion)</td>
<td>26.8</td>
<td>19.3</td>
<td>35.2</td>
<td>32.6</td>
<td>55.3</td>
</tr>
<tr>
<td>Eigen et al. (Coarse)</td>
<td>27.9</td>
<td>23.4</td>
<td>34.5</td>
<td>25.5</td>
<td>48.4</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>RMSE 11.25°</td>
<td>RMSE 22.5°</td>
<td>RMSE 30°</td>
</tr>
<tr>
<td>----------------</td>
<td>------</td>
<td>--------</td>
<td>-------------</td>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td>Full</td>
<td>25.0</td>
<td>13.8</td>
<td>35.9</td>
<td>44.2</td>
<td>63.2</td>
</tr>
<tr>
<td>Full (Soft)</td>
<td>24.2</td>
<td>17.3</td>
<td>32.2</td>
<td>36.8</td>
<td>58.5</td>
</tr>
<tr>
<td>Fusion (+VP)</td>
<td>25.3</td>
<td>14.4</td>
<td>35.9</td>
<td>42.7</td>
<td>62.5</td>
</tr>
<tr>
<td>Fusion (+Edge)</td>
<td>25.8</td>
<td>15.3</td>
<td>36.0</td>
<td>40.0</td>
<td>61.6</td>
</tr>
<tr>
<td>Fusion (+Layout)</td>
<td>25.8</td>
<td>14.9</td>
<td>36.3</td>
<td>41.1</td>
<td>61.9</td>
</tr>
<tr>
<td>Fusion</td>
<td>26.0</td>
<td>15.5</td>
<td>36.2</td>
<td>39.5</td>
<td>61.3</td>
</tr>
<tr>
<td>Bottom-up</td>
<td>32.2</td>
<td>23.5</td>
<td>42.0</td>
<td>27.2</td>
<td>48.5</td>
</tr>
<tr>
<td>Top-down</td>
<td>29.0</td>
<td>19.8</td>
<td>38.3</td>
<td>32.7</td>
<td>53.8</td>
</tr>
</tbody>
</table>

Eigen et al. (Fusion) | 26.8 | 19.3 | 35.2 | 32.6 | 55.3 | 65.5 |
Eigen et al. (Coarse) | 27.9 | 23.4 | 34.5 | 25.5 | 48.4 | 60.6 |
Thoughts
Do not address bas-relief
Incorporate Computer Graphics
Inverse problem
Given surface normals
How should the scene look?
What is the correct image?
Incorporate image formation model
Why depth from single image