Siamese Network & Stereo

Wenjie Luo
CSC2523
Feb 2nd, 2016
Outline

• Siamese network
• Application: stereo
• Discussion
• Recap on CNN:
  - Input: one image
  - Output: class label, bounding box etc..

• What if?
  - Input: two images, equivalent
  - Parameter sharing?
Siamese network

• Consists of two identical sub-networks: feature extraction
• Joined at their outputs: measure distance between feature vectors
• Date back to NIPS 1994

Source: J. Bromley et. al.
Applications

• Face verification/recognition
• Video sequence
• *Stereo* (depth estimation)
Why depth

• Structure and depth are inherently ambiguous from a single view
Stereo

- Estimate depth from stereo images.

- Depth is inversely proportional to disparity.

$$Z = f \frac{B}{d}$$

Z: depth; f: focal length; B: baseline; d: disparity
We need..

- Correspondances on image locations (Matching)
  - *Good feature*
- Refinement in practice
  - Smoothing
Conv-Nets

• Input: two image patches
  - Equivalent

• Output: matching cost

• What architecture would you use?
Network I

- Two stages:
  - Siamese network
  - Fully connected
- Input: small patch
- Binary prediction
- “Big” network (~600K)

Source: Zbontar & LeCun
Network II

- Dot-product
- Input: full content
- Larger patch
- Log loss
- Smaller network
- Gray image, outdoor/noisy, 194/195 split
- Disparity range: 256
- Saturation/Textureless(dynamic range)
- Evaluation metric
Training

• Preprocessing
  - full image or small patch
  - data-augmentation, loading

• Siamese network
  - Gradient aggregated

• Initialization, SGD

• Batch Normalization(variance shift, works well)
Test

- Image size: W, H; Disparity range: D
  \[ W \times H \times D = 1200 \times 370 \times 256 = 1.14 \times 10^8! \]

- Computation
  - Feature shared

- Memory
  - One disparity at a time
Smoothing

- Cost-aggregation
  - Averaging neighboring locations

- CRF
  - Semiglobal matching

- Post-processing
  - Border fixing(CNN), left-right consistency, outlier detector
## Stereo Evaluation 2012

<table>
<thead>
<tr>
<th>Table</th>
<th>All</th>
<th>Error threshold</th>
<th>Evaluation area</th>
<th>All pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Setting</td>
<td>Code</td>
<td>Out-Noc</td>
<td>Out-All</td>
</tr>
<tr>
<td>1</td>
<td>Displets v2</td>
<td>code</td>
<td>2.37 %</td>
<td>3.09 %</td>
</tr>
<tr>
<td>2</td>
<td>MC-CNN-acrt</td>
<td>code</td>
<td>2.43 %</td>
<td>3.63 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>J. Zbontar and Y. LeCun: <strong>Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches</strong>. Submitted to JMLR.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Displets</td>
<td>code</td>
<td>2.47 %</td>
<td>3.27 %</td>
</tr>
<tr>
<td>4</td>
<td>MC-CNN</td>
<td></td>
<td>2.61 %</td>
<td>3.84 %</td>
</tr>
<tr>
<td>5</td>
<td>PRSM</td>
<td>code</td>
<td>2.78 %</td>
<td>3.00 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. Vogel, K. Schindler and S. Roth: <strong>3D Scene Flow Estimation with a Piecewise Rigid Scene Model</strong>. ijcv 2015.</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>SPS-STFL</td>
<td></td>
<td>2.83 %</td>
<td>3.64 %</td>
</tr>
<tr>
<td>7</td>
<td>VC-SF</td>
<td></td>
<td>3.05 %</td>
<td>3.31 %</td>
</tr>
<tr>
<td>8</td>
<td>Deep Embed</td>
<td></td>
<td>3.10 %</td>
<td>4.24 %</td>
</tr>
<tr>
<td>9</td>
<td>JSOM</td>
<td></td>
<td>3.15 %</td>
<td>3.94 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Anonymous submission</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>OSF</td>
<td>code</td>
<td>3.28 %</td>
<td>4.07 %</td>
</tr>
</tbody>
</table>
Thank You

Q&A