Hand Pose Estimation through Weakly-Supervised Learning of a Rich Intermediate Representation

Natalia Neverova  Christian Wolf  Graham W. Taylor  Florian Nebout
Deep Learning of Human Motion

Gesture recognition

Recognition of individual activities

Pose estimation
Hand pose estimation: problem formulation
Dataset 1: real data (NYU Hand Pose Dataset)
Dataset 2: synthetic data (ours)
Why an intermediate representation?

- Contains rich structural and topological information, since the label space itself is structured.
- Labels have adjacency, topological and geometric relationships.
- Can be leveraged and translated into loss for weakly supervised training;
- Eases joint regression when combined with raw input
Why an intermediate representation?

Dense segmentation is richer than joint positions and can’t be recovered from them.
Proposed method

CNN

depth map

weakly supervised learning

intermediate representation

joint coordinates
Functional overview

- Segmentation learner $f_s$
  - Produced segmentation maps
  - Synthetic data: segmentation maps
  - Real data: depth images
- Dictionary of synthetic patches
- Patchwise restoration
- Error segmentation vs joints
- Regression learner $f_r$
  - Produced joint coordinates
- Real data labels: joint coordinates
- Real data: depth images
- Region localization
segmentation learner $f_s$

synthetic data: segmentation maps

produced segmentation maps

synthetic data: depth images

region localization

error segmentation vs joints

dictionary of synthetic patches

patchwise restoration

real data labels: joint coordinates

produced joint coordinates

real data: depth images

regression learner $f_r$
segmentation learner $f_s$

synthetic data: segmentation maps

produced segmentation maps

error segmentation vs joints

dictionary of synthetic patches

patchwise restoration

regression learner $f_r$

real data labels: joint coordinates

produced joint coordinates

region localization

real data: depth images

synthetic data: depth images
Learning process

- **synthetic data:** segmentation maps
- **real data:** depth images

**Paths:**
- **Learning process**
- **segmentation learner** $f_s$
- **dictionary of synthetic patches**
- **patchwise restoration**
- **error segmentation vs joints**
- **regression learner** $f_r$
- **region localization**
- **real data labels:** joint coordinates
- **produced joint coordinates**

**Produced data:**
- segmentation maps
- depth images

**Real data:**
- depth images
- joint coordinates

**Weakly supervised training:**
- $f_s$ (synthetic data)
- $f_r$ (real data)
synthetic data: segmentation maps

produced segmentation maps

dictionary of synthetic patches

patchwise restoration

error segm. vs joints

regression learner $f_r$

produced joint coordinates

real data labels: joint coordinates

region localization

real data: depth images

synthetic data: depth images
Architecture: regression learner
Regression learner $f_r$

Produced segmentation maps

Synthetic data: depth images

Segmentation learner $f_s$

Error segmentation vs joints

Patchwise restoration

Dictionary of synthetic patches

Real data labels: joint coordinates

Real data: depth images

Region localization

Produced joint coordinates

Synthetic data: segmentation maps
Restoration of segmentation maps by kNN search
Restoration of segmentation maps by approximate kNN search

Center pixel restoration

$$\nu(q^{(j)}) = \arg \min_{p^{(i)} \in P} d_H(q^{(j)}, p^{(l)})$$

Integration over overlapping patches

$$f_{nn}(q^{(j)}) = \arg \max_l \sum_{k \in W \times W} \mathbb{I}(l = n^{(k,j \oplus k)})$$
Restoration of segmentation maps by kNN search
### Experimental results: restoration of segmentation maps

<table>
<thead>
<tr>
<th>Method</th>
<th>— per pixel —</th>
<th>— per class —</th>
</tr>
</thead>
<tbody>
<tr>
<td>No restoration</td>
<td>51.03</td>
<td>39.38</td>
</tr>
<tr>
<td>NN-search — no integration</td>
<td>48.76</td>
<td>39.72</td>
</tr>
<tr>
<td><strong>NN-search — integration</strong></td>
<td><strong>54.55 (+3.52)</strong></td>
<td><strong>46.38 (+7.00)</strong></td>
</tr>
<tr>
<td>CRF – Potts-like model</td>
<td>53.10 (+2.07)</td>
<td>43.64 (+4.26)</td>
</tr>
<tr>
<td>CRF – Hamming distance on overlapping patch area</td>
<td>52.45 (+1.42)</td>
<td>42.68 (+3.30)</td>
</tr>
<tr>
<td>Method</td>
<td>Datasets used</td>
<td>— per pixel—</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Fully supervised training only</td>
<td>synth. segmentations</td>
<td>51.03</td>
</tr>
<tr>
<td>Semi-/weakly-supervised training</td>
<td>synth. segmentations + real joint positions</td>
<td><strong>57.18</strong> (+6.15)</td>
</tr>
</tbody>
</table>

Experimental results: weakly supervised learning
<table>
<thead>
<tr>
<th>Method</th>
<th>Mean 2D, mm</th>
<th>Median 2D, mm</th>
<th>Mean 3D, mm</th>
<th>Median 3D, mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct regression</td>
<td>13.27</td>
<td>10.56</td>
<td>17.26</td>
<td>14.45</td>
</tr>
<tr>
<td>Cascade direct regression, 1 step</td>
<td>12.56</td>
<td>9.94</td>
<td>16.88</td>
<td>14.10</td>
</tr>
<tr>
<td>Regression incl. segmentation</td>
<td><strong>11.18</strong></td>
<td><strong>8.67</strong></td>
<td><strong>14.94</strong></td>
<td><strong>13.62</strong></td>
</tr>
<tr>
<td>Multi-Scale, Oberweger et al. [14]</td>
<td>-</td>
<td>-</td>
<td>27.5[^1]</td>
<td>-</td>
</tr>
<tr>
<td>Deep, Oberweger et al. [14]</td>
<td>-</td>
<td>-</td>
<td>30.5[^1]</td>
<td>-</td>
</tr>
<tr>
<td>Shallow, Oberweger et al. [14]</td>
<td>-</td>
<td>-</td>
<td>34.5[^1]</td>
<td>-</td>
</tr>
<tr>
<td>Tompson et al. [32]</td>
<td>7.05</td>
<td>6.54</td>
<td>21.0[^1][14]</td>
<td>-</td>
</tr>
<tr>
<td>Xu and Cheng [34]</td>
<td>-</td>
<td>58.0[^1][8]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Keskin et al. [11]</td>
<td>-</td>
<td>72.5[^1][8]</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Experimental results: estimation of joint locations
Conclusion

- Hand pose estimation from row depth input
- Intermediate representation as latent variable during training
- Weakly- & semi-supervised learning from simulated and real data
- Structural and topological properties can be leveraged to create loss during training