Nonparametric Structure Regularization Machine for 2D Hand Pose Estimation

Yifei Chen*, Haoyu Ma*, Deying Kong², Xiangyi Yan², Jianbo Wu¹, Wei Fan¹, and Xiaohui Xie²
¹ Tencent Hippocrates Research Lab
² Department of Computer Science, University of California at Irvine

Introduction

- Objective: 2D hand pose estimation (keypoint detection)
- Application: AR/VR, gesture recognition, basic for 3D task.
- Challenge: self-occlusion due to articulation, viewpoint and object.
- Current Approach:
  - Deep convolutional neural network: Convolutional Pose Machines (CPM) and Stacked Hourglass, only capturing pose structure information implicitly.
  - Multi-task learning: unify hand pose estimation with hand mask segmentation, requiring a large amount of manually labelled mask for hand.
- Our Contributions:
  - We propose a novel cascade structure regularization methodology for 2D hand pose estimation, which utilizes synthetic hand masks to guide keypoints structure learning.
  - We propose a novel probabilistic representation of hand limbs and an anatomically inspired composition strategy for hand mask synthesis.

Learning

- Loss: \( \text{Loss} = \text{Loss}_{\text{keypoint}} + \lambda_1 \text{Loss}_{\text{structure}}^{G1} + \lambda_2 \text{Loss}_{\text{structure}}^{G6} \)
- Training Strategy:
  - End-to-End Training
  - Decayed loss schedule: Structure learning is an auxiliary task, thus there is no need to get an accurate results, and our ultimate goal is keypoint. Let \( \lambda_1 \) and \( \lambda_2 \) decay by a ratio of 0.1 every 20 epochs during training.

Contact: haoyum3@uci.edu
Code: https://github.com/HowieMa/NSRMhand

Methodology

- Limb Mask Representation: Generate synthetic limb mask from labeled keypoints
  - Hand model: 21 Keypoints + 20 Limbs L (Line Segment)
  \( S_{\text{LDM}}(p|L) = \begin{cases} 1 & \text{if } p \in L \\ 0 & \text{otherwise} \end{cases} \)
  \( S_{\text{LPM}}(p|L) = \exp(-\frac{\|p - P_{\text{CPM}}(L)\|^2}{2\sigma^2}) \)
- Limb Composition
  - G1: coalesce 20 limbs together (whole hand mask)
  - G6: coalesce 20 limbs into 6 groups (5 fingers + palm)
  \( S = \max(S_{\text{LMP}}(p|L)) \)

In practice, we mainly focus on utilizing G1 and G1&6 (the combination of G1 and G6).

Network Architecture: based on CPM

Results

- Quantitative Results:
  - Probability of Correct Keypoint (PCK) curve on Onehand10k and panoptic hand dataset
  - PCK value on Panoptic dataset

<table>
<thead>
<tr>
<th>PCK</th>
<th>0.99</th>
<th>0.98</th>
<th>0.97</th>
<th>0.96</th>
<th>0.95</th>
<th>0.94</th>
<th>0.93</th>
<th>0.92</th>
<th>0.91</th>
<th>0.90</th>
<th>0.89</th>
<th>0.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPM</td>
<td>55.25</td>
<td>73.23</td>
<td>81.45</td>
<td>85.97</td>
<td>88.80</td>
<td>76.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDM-G1</td>
<td>59.20</td>
<td>75.98</td>
<td>83.45</td>
<td>87.28</td>
<td>89.81</td>
<td>79.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDM-G1&amp;6</td>
<td>59.16</td>
<td>76.32</td>
<td>83.63</td>
<td>87.46</td>
<td>90.03</td>
<td>79.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPM-G1</td>
<td>59.81</td>
<td>76.82</td>
<td>84.16</td>
<td>87.86</td>
<td>90.26</td>
<td>79.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPM-G1&amp;6</td>
<td>59.73</td>
<td>76.86</td>
<td>84.43</td>
<td>88.23</td>
<td>90.87</td>
<td>80.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Qualitative Results

CPM

NSRM