Nonparametric Structure Regularization Machine for 2D Hand Pose Estimation

Yifei Chen∗1, Haoyu Ma∗2, Deying Kong2, Xiangyi Yan2, Jianbao Wu1, Wei Fan†1, and Xiaohui Xie†2
1Tencent Hippocrates Research Lab
2Department of Computer Science, University of California at Irvine
{dolphinchen,jannwu,davidwfan}@tencent.com, {haoyum3,deyingk,x.yan,xhx}@uci.edu

Abstract

Hand pose estimation is more challenging than body pose estimation due to severe articulation, self-occlusion and high dexterity of the hand. Current approaches often rely on a popular body pose algorithm, such as the Convolutional Pose Machine (CPM), to learn 2D keypoint features. These algorithms cannot adequately address the unique challenges of hand pose estimation, because they are trained solely based on keypoint positions without seeking to explicitly model structural relationship between them. We propose a novel Nonparametric Structure Regularization Machine (NSRM) for 2D hand pose estimation, adopting a cascade multi-task architecture to learn hand structure and keypoint representations jointly. The structure learning is guided by synthetic hand mask representations, which are directly computed from keypoint positions, and is further strengthened by a novel probabilistic representation of hand limbs and an anatomically inspired composition strategy of mask synthesis. We conduct extensive studies on two public datasets - OneHand 10k and CMU Panoptic Hand. Experimental results demonstrate that explicitly enforcing structure learning consistently improves pose estimation accuracy of CPM baseline models, by 1.17% on the first dataset and 4.01% on the second one. The implementation and experiment code is freely available online.

1. Introduction

Hand pose understanding is an important task for many real world AI applications, such as human-computer interaction, augmented reality and virtual reality. However, hand pose estimation remains challenging because the hand is highly articulated and dexterous, and suffers severely from self-occlusion. Recently a significant amount of efforts have been dedicated to improving the accuracy of hand pose estimation from different perspectives, including 1) multi-view RGB systems [23, 11], 2) depth-based solutions [7, 28, 32], and 3) monocular RGB solutions [35, 19, 2]. Although some of these efforts focus on 3D hand pose or shape estimation, 2D hand pose estimation remains an essential component as it often constitutes a sub-module of 3D estimation problems and as such directly impacts the performance of downstream 3D pose or shape estimation.

Meanwhile, human pose estimation has advanced significantly since the advent of Deep Convolutional Neural Network (DCNN). Successful DCNN architectures typically have large receptive fields and strong representation power, such as the Convolutional Pose Machine (CPM) [30], the Stacked Hourglass (SHG) [18], and the Residual Network [10]. They are deployed by popular human pose estimation systems [3, 6, 9] to implicitly capture structure information of body parts. They are also utilized by many hand pose estimation algorithms to perform the 2D pose estimation subtask [23, 35, 2, 19, 28]. However, DCNNs only capture structure information implicitly and may not be adequately equipped to capture complex structure relationship between hand keypoints to handle severe articulation and self-occlusion of the hand [12, 14].

Recently, there is a trend to unify pose estimation and instance segmentation in a multi-task learning paradigm, and it is observed that the latter helps to improve the performance of the former [9, 29]. Unfortunately, this direction requires a large amount of manually labelled segmentation masks, which is costly to obtain. Hand mask datasets are even rarer than body mask datasets, making the multi-task approach less applicable to hand pose estimation.

In this paper, we propose the Nonparametric Structure Regularization Machine (NSRM) for 2D hand pose esti-
mation from a monocular RGB image. NSRM incorporates a nonparametric structure model and a pose model in a cascade multi-task framework. The structure learning is supervised by synthetic hand mask representations directly computed from keypoint positions, and is strengthened by a probabilistic representation of hand limbs and an anatomically inspired composition strategy of mask synthesis. The pose model utilizes the composite structure representation to learn robust hand pose representation. We comprehensively evaluate the performance of NSRM on two public datasets, i.e., OneHand 10k [29], and the more challenging CMU Panoptic Hand [11]. Quantitative results demonstrate that NSRM consistently improves the prediction accuracy of the CPM baseline, by 1.17% on the first dataset and 4.01% on the second one, and that NSRM renders competitive performance compared to utilizing manually-labeled masks. Qualitative results show that NSRM effectively reinforces structure consistency to predicted hand pose especially when severe occlusion exists, and the learned structure representations highly resemble real segmentation masks.

The main contributions of this paper are as follows:

- We propose a novel cascade structure regularization methodology for 2D hand pose estimation, which utilizes synthetic hand masks to guide keypoints structure learning. The synthetic masks are derived directly from keypoint positions requiring no extra data annotations, making the method applicable to any existing pose estimation model.

- We propose a novel probabilistic representation of hand limbs and an anatomically inspired composition strategy for hand mask synthesis.

- We carry out extensive experiments on two public datasets, and demonstrate that NSRM consistently outperforms baseline models.

2. Related work

2.1. Human pose estimation

DCNN has been massively applied to 2D human pose estimation since the seminal work of DeepPose [27]. As the human body naturally manifests an articulated graph structure, researchers have explored the combination of DCNN and graphical models (GM) for pose estimation [26, 4, 31, 25]. However, the GM component often suffers from two practical limitations: 1) the pairwise term typically takes some parametric form, which may not be true in reality; 2) belief propagation inference is performed frequently during training and computational intensive. As a result, mainstream algorithms [3, 6, 9] still rely on delicate DCNN architectures, such as CPM [30], SHG [18], and the Residual Network [10] to implicitly capture structure information, deploying their large receptive fields and strong representation power. To further improve the effect of structure regularization, some approaches attempt to modify the output of DCNN, for example, to introduce extra branches of the offset field [20] or the structural-aware loss [12]. Unfortunately, they still cannot fully characterize the structure conditions of limbs, especially their poses and interactions, resulting in very limited effect.

Meanwhile, 3D human pose estimation from monocular RGB, a challenger problem due to depth ambiguity, also advances significantly. Some researches explicitly infer 3D coordinates from the 2D pose [1, 15, 22, 21], while others incorporate 2D pose estimation networks into the whole architecture [16, 17, 33]. In both cases, DCNN-based 2D pose estimators are intensively utilized, such as the Mask RCNN [9] and SHG. To enforce structure constraints in 3D, a kinematic layer can be added on top of the network [34]. But this relies on known bone length and may suffer from the optimization difficulty, which limit its practical application.

2.2. Hand pose estimation

Hand pose estimation is more difficult than body pose estimation, as the hand is highly articulated, dexterous, and suffers severely from self-occlusion. Although multi-view camera systems can solve the task reliably [23, 11], it is usually very costly to build such a system, involving numerous optical devices and complicated configuration. Therefore, their applications are mostly restricted to the laboratory scenarios. To circumvent this limitation, researchers also devote much effort to depth-based solutions [7, 28, 32]. However, depth devices have limited resolution and range, and are sensitive to lighting conditions [17]. And after all, they are still less ubiquitous than RGB cameras.

Due to the drawbacks of multi-view and depth-based solutions, monocular RGB approaches are drawing much more attention in recent years. Like in the case of 3D human pose estimation, many algorithms adopt a two-stage framework, i.e., first performing 2D hand pose estimation and then lifting from 2D to 3D [35, 19, 2]. The 2D subtasks commonly utilize prevalent 2D human pose algorithms, in particular CPM and SHG, which are also frequently used in the multi-view or depth-based solutions [23, 28]. Given the critical role of 2D hand pose algorithms in solving the complete 3D task, we focus on imposing novel structure regularization to 2D hand pose estimation in this paper.

3. The model

Nonparametric Structure Regularization Machine (NSRM) learns the hand’s structure and keypoint representations in a cascade multi-task framework. A high-level illustration of the hand representation and our overall architecture is shown in Figure 1. The hand is modeled as
21 keypoints and 20 limbs. The former are interconnected via the latter anatomically. First, the backbone module, usually a DCNN, e.g., the VGG [24], processes an input image and generates a set of feature maps. Second, the structure module takes in image feature maps to learn the mask representation of the hand. Third, the pose module takes in both the mask representation and the feature maps to learn the pose representation of the hand, i.e., keypoint confidence maps (KCM). Both the structure and the pose modules are multi-stage convolutional neural networks.

3.1. Limb mask representation

Consider any particular limb \( L \) between Keypoint \( i \) and \( j \) as defined in Figure 1a. We define our basic mask representation as follows,

**Limb Deterministic Mask (LDM).** Pixels that belong to \( L \) are defined as those which fall inside a fixed-width rectangle centering around line segment \( \overline{p_ip_j} \), i.e.,

\[
\begin{align*}
0 \leq (p - p_j)^T (p_i - p_j) & \leq \|p_i - p_j\|^2, \\
(p - p_j)^T u & \leq \sigma_{LDM}
\end{align*}
\]

where \( u \) is a unit vector perpendicular to \( \overline{p_ip_j} \), and \( \sigma_{LDM} \) is a hyper parameter to control the width of the limb. The ground truth of LDM is a simple 0/1-mask defined outside/inside the rectangle, i.e.,

\[
S_{LDM}(p|L) = \begin{cases} 
1 & \text{if } p \in L \\
0 & \text{otherwise}
\end{cases}
\]

where \( p \in I \) is an arbitrary pixel in the image. See Figure 2a for an illustration.

**Limb Probabilistic Mask (LPM).** LDM assigns 0/1 value to each pixel, depending on its belong to the rectangular mask. This crude treatment may not be optimal in practice. We further propose the novel LPM representation. Each pixel belongs to \( L \) with a Gaussian-alike confidence value as defined bellow,

\[
S_{LPM}(p|L) = \exp \left( - \frac{D(p,p_ip_j)^2}{2\sigma_{LPM}^2} \right)
\]

where \( D(p,p_ip_j) \) is the distance between the pixel \( p \) and the line segment \( \overline{p_ip_j} \), and \( \sigma_{LPM} \) is a hyper parameter to control the spread of the Gaussian. See Figure 2b for an illustration. LPM is a smoothed expansion of LDM.
3.2. Limb composition

Given mask representations of single limbs, we further coalesce them into anatomically legitimate groups. Our basic strategy is to coalesce all the 20 limbs together, which renders one single mask representing the whole hand (denoted as $G1$). Alternatively, we also consider coalescing limbs separately into six groups, one representing each finger and one representing the palm (denoted as $G6$). $G1$ captures the overall structure of the hand, while $G6$ concerns more about detailed structure regarding local areas of the hand. See Figure 3 and 4 for an illustration. Formally, consider any particular limb group $g$ containing $|g|$ limbs, $\{L_1, L_2, ..., L_{|g|}\}$. Using limb composition, the coalesced mask is defined as,

$$S^*(p|g) = \max \{S(p|L_1), S(p|L_2), ..., S(p|L_{|g|})\}$$ (4)

where $S(p|L)$ is computed using the basic representation of $S_{LDM}$ (Equation 2) or $S_{LPM}$ (Equation 3).

In practice, we mainly focus on utilizing $G1$ and $G1&6$ (the combination of $G1$ and $G6$). We note that although $G1$ resembles the hand segmentation mask, it is much more efficient because it is readily obtained from keypoints without the extra work of mask annotation. In Section 4, we will compare the performance of utilizing our LDM/LPM representations against the real segmentation mask.

Figure 3: The $G1$ composition strategy: coalescing all the 20 limbs together to get one single mask representing the whole hand.

3.3. Loss function and training

Intermediate supervision is applied to each stage of the structure module and the pose module. Following the common practice of instance segmentation [9], we apply cross-entropy loss to the output of our structure module, i.e.,

$$L_S = \sum_{t=1}^{T_S} \sum_{g \in G} \sum_{p \in I} S^*(p|g) \log \hat{S}_t(p|g)$$

$$+ (1 - S^*(p|g)) \log \left(1 - \hat{S}_t(p|g)\right)$$ (5)

where $T_S$ is the number of stages of structure learning, and $\hat{S}_t(p|g)$ is the prediction of the structure module at Pixel $p$, Group $g$ of Stage $t$.

Following the common practice of pose estimation [18, 3], we define the ground truth KCM of Keypoint $k$ as a 2D Gaussian centering around the labelled keypoint with standard deviation $\sigma_{KCM}$, i.e.,

$$C^*(p|k) = \exp \left\{ -\frac{\|p - p_k^C\|^2}{2\sigma_{KCM}^2} \right\}$$ (6)

We apply the sum-of-squared-error loss to the output of our pose module, i.e.,

$$L_K = \sum_{t=1}^{T_K} \sum_{k=1}^{K} \sum_{p \in I} \left\|C^*(p|k) - \hat{C}_t(p|k)\right\|^2_2$$ (7)

where $T_K$ is the number of stages of pose learning, and $\hat{C}_t(p|k)$ is the prediction of the pose module at Pixel $p$, Keypoint $k$ of Stage $t$.

The overall loss function is thus a weighted sum of the structure loss and the pose loss, i.e.,

$$L = \begin{cases} L_K + \lambda_1 L_S^G1, & \text{for } G1 \\ L_K + \lambda_1 L_S^G1 + \lambda_2 L_S^G6, & \text{for } G1&6 \end{cases}$$ (8)

where $\lambda_1, \lambda_2$ are hyper-parameters to control the relative weight of the structural regularization. The whole system is trained end-to-end.

4. Experiments

4.1. Datasets

We evaluate NSRM on two public hand pose datasets: the OneHand 10k dataset [29] (OneHand 10k), and the
CMU Panoptic Hand dataset [11] (Panoptic). Their overall statistics are summarized in Table 1. More descriptions are as follows.

**OneHand 10k** contains 11,703 in-the-wild hand images annotated with both segmentation masks and keypoints. Being collected online, the images often have cluttered background and cover various hand poses. Invisible keypoints are often left unannotated. Ground truth limb representations related to these missing keypoints are set to zero maps. We don’t do any hand cropping, as most of the images are occupied by one hand. The dataset is already partitioned into training and testing subsets by Wang et al. [29].

**Panoptic** contains 14,817 images of persons from the Panoptic Studio, each with 21 annotated keypoints of the right hand. Since we focus on hand pose estimation instead of hand pose detection, we directly crop hands based upon their ground truth keypoints. Specifically, we crop a square patch of size 2.2B, where B is the maximum dimension of the tightest bounding box enclosing all hand keypoints. Then the cropped hand dataset is randomly divided into three subsets for training (80%), validation (10%) and testing (10%).

Table 1: Overall statistics of datasets used in this paper.

<table>
<thead>
<tr>
<th>dataset</th>
<th>training</th>
<th>validation</th>
<th>testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>OneHand 10k</td>
<td>10,000</td>
<td>-</td>
<td>1,703</td>
</tr>
<tr>
<td>Panoptic</td>
<td>11,853</td>
<td>1,482</td>
<td>1,482</td>
</tr>
</tbody>
</table>

### 4.2. Experimental settings

**Implementation details**

We implement NSRM in Pytorch. To be compatible with and comparable to the CPM hand pose model [23], we adopt VGG-19 [24] (up to Conv 4,4) as our backbone network, which is pretrained on ImageNet [5] and generates 128-channel feature maps. The following architecture has 6 stages, each of which contains 5 convolution layers with 7x7 kernel and 2 convolution layers with 1x1 kernel (except the first stage). The first 3 stages are allocated to learn composite mask representations, and the last 3 stages are for pose representation learning. All hand image patches are resized to 368 × 368 before fed into our model, yielding 46 × 46 representation maps for both LDM/LPM and KCM. The detail network architecture is shown in Figure 5.

**Learning configuration**

We use Adam [13] to train our model. The initial learning rate is set to 1e-4, and other parameters are set to default values. For G1, we set λ1 = 1 for LDM, and λ1 = 0.5 for LPM. For G1&6, we set λ1 = 0.2, λ2 = 0.04 for LDM, and λ1 = 0.1, λ2 = 0.02 for LPM. These configurations empirically make the structure loss and the pose loss on the same scale at the beginning of training. Further more, as structure learning is an auxiliary task and our ultimate goal is hand pose estimation, we propose to utilize a decayed loss training schedule. Specifically, we let λ1 and λ2 decay by a ratio of 0.1 every 20 epochs, so as to let training focus more on KCM in later iterations.

**Evaluation metric**

We adopt Probability of Correct Keypoint within a Normalized Distance Threshold [23] (shortly referred to as PCK) to perform model selection and evaluation. However, as the hand/head sizes are not explicitly provided by the datasets used in this paper, we resort to normalization with respect to the dimension of the tightest hand bounding box. The normalization threshold σ_{PCK} ranges from 0 to 1.

### 4.3. Quantitative results

**OneHand 10k**

As this dataset has a lot of missing values, G6 composition tends to generate incomplete masks. Therefore, we only consider about G1 composition. We retrain the Maskpose Cascaded CNN [29], which utilize real segmentation masks (denoted as “Real Mask”). We also train the model with our proposed decayed loss training schedule (denoted as “Real Mask ++”). Figure 6a shows the performance comparison. Table 2 summarizes detailed numerical results. Our observations are as follows:

i) NSRM consistently improves the predictive accuracy of CPM, regardless of the choice of basic representation (LDM or LPM), and the value of the evaluation threshold σ_{PCK}. In particular, LPM-G1 achieves 0.0102 absolute improvement in average PCK, corresponding to 1.17% relative improvement.

ii) LPM-G1 outperforms LDM-G1, and does slightly better than Real Mask. This result demonstrates the effectiveness of our proposed probabilistic mask, comparing to both the manually labeled mask and our proposed deterministic mask.

iii) Our proposed decayed loss training schedule, in combination with utilizing real masks, achieves the best performance (Real Mask ++).

To summarize, our proposed NSRM (along with its learning schedule) is both efficient and effective, as it avoids the overhead of mask labeling but still maintains competitive performance.

**Panoptic**

Figure 6b shows the performance of NSRM and the CPM baseline. Table 3 summarizes detailed numerical results. Our observations are as follows:

i) Like in the case of OneHand 10k, NSRM consistently outperforms the CPM baseline. In particular, the fully-fledged LPM-G1&6 achieves 0.0309 absolute improvement in average PCK, corresponding to 4.01% relative improvement. The results suggest a systematic and significant ben-
ii) Like in the case of OneHand 10k, LPM consistently outperforms LDM, under both composition strategies (G1 and G1&6). The absolute improvement in average PCK is 0.64 under G1 and 0.71 under G1&6. This phenomenon consolidates that the probabilistic treatment in mask representation indeed benefits the performance of NSRM.

iii) Comparing the proposed composition strategies, we find that G1&6 moderately improves the performance of G1. We interpret this result from an anatomical perspective. As an overall representation, G1 covers important global structure information of the hand. However it cannot fully characterize local details, such as the shape of each finger, which are highly flexible and hard to distinguish due to self-occlusion. G6 is designed to cope with this problem. By combining G1 and G6, NSRM gets a representation that can cover both global and local structure information.

iv) The combination of centralized & distributed composition and the probabilistic representation (LPM-G1&6) renders the optimal performance, which correspond to 0.89 absolute improvement in average PCK, comparing to the basic version (LDM-G1).

v) Although being consistent on both datasets, the improvement seems much more significant on Panoptic than on OneHand 10k. There are two potential reasons. First, Panoptic is much more challenging, with abundant hand gestures shot from different perspectives and complicated hand-hand & hand-object interactions. Second, OneHand 10K contains a lot of unlabelled keypoints, and therefore tends to make the mask learning signal too fragmented and noisy. Both factors suggest that Panoptic could benefit more from structure learning than OneHand 10k.

4.4. Qualitative results

Figure 7 visualizes the predictive results of CPM and NSRM on sample images from the Panoptic test data. We can clearly see that NSRM effectively reinforces structure
consistency and reduces inference ambiguity. Even during severe occlusion, LPM-G1&6 still makes anatomically legitimate prediction while CPM cannot. Moreover, Figure 8 visualizes the learned structure representations on sample images. We can see that they highly resemble hand segmentation masks. This indicates that our NSRM framework could be potentially applied to multi-task learning of both hand pose estimation and instance segmentation.

4.5. Discussion

NSRM learns compositional structure representation of the hand, and utilizes it to regularize the learning process of KCM in a nonparametric fashion. This implicit hierarchical treatment is fundamentally different to graphical-model-involved approaches [4, 31, 25, 14], and those which introduce simple keypoint-induced output/loss [20, 12] or kinematic constraints [34].

Our basic LDM representation takes a rectangular shape, similar to the Part Affinity Field (PAF) [3]. However, NSRM goes significantly beyond PAF in three core aspects. First, our segmentation-inspired representation and its probabilistic expansion are completely different from the vector field representation of PAF. Second, we propose structure composition to coalescing limbs into anatomically inspired groups, a key feature not considered by PAF. Last but not least, PAF is an auxiliary representation proposed for differentiating multiple human instances, not intended for hand structure regularization as in our case.

Previous researches have explored simultaneous pose estimation and instance segmentation [9, 29], but all require mask annotation. Our structure representation is automatically constructed from keypoints, avoiding laborious anno-

Table 3: Detailed numerical results of PCK (in %) evaluated at different thresholds on the Panoptic testing data. “ave” means the average PCK, whose absolute and relative improvement are shown in the right most column. The best improvement is highlighted in boldface.

<table>
<thead>
<tr>
<th>$\sigma_{PCK}$</th>
<th>0.04</th>
<th>0.06</th>
<th>0.08</th>
<th>0.10</th>
<th>0.12</th>
<th>ave</th>
<th>improvement</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>
</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>+2.20 (+2.86%)</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>+2.38 (+3.09%)</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>+2.84 (+3.69%)</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>+3.09 (+4.01%)</td>
</tr>
</tbody>
</table>

Figure 6: PCK curves on testing data of a) OneHand 10k, and b) Panoptic. Best viewed in color.
tation. Our experiments demonstrate that NSRM achieves comparable pose estimation performance to models trained with real masks [29]. Moreover, our learned structure representations highly resemble real masks, which indicates potential applications to hand instance segmentation.

5. Conclusion

In this paper, we have proposed a novel Nonparametric Structure Regularization Machine for 2D hand pose estimation. NSRM is a cascade architecture of structure learning and pose learning. The structure learning is guided by self-organized hand mask representations, and strengthened by a novel probabilistic representation of hand limbs and an anatomically inspired composition strategy of mask synthesis. The pose module utilizes the structure representation to learn robust hand pose representation. We comprehensively evaluate NSRM on two public datasets. Experimental results demonstrate that, 1) NSRM consistently improves the prediction accuracy of the baseline model; 2) NSRM renders comparable performance to utilizing manually labeled masks; 3) NSRM effectively reinforces structure consistency to predicted hand poses, especially during severe occlusion. We should note that although we used CPM as our baseline pose estimation model, the proposed method is generic, independent of any particular choice of the baseline. The structure learning module can be readily added to other prevalent models [18, 8] to improve pose estimation performance, which we intend to explore in future work.
References


