HandMap: Robust Hand Pose Estimation via Intermediate Dense Guidance Map Supervision

Xiaokun Wu, Daniel Finnegan, Eamonn O'Neill and Yong-Liang Yang

Department of Computer Science, University of Bath, UK

{xw943,d.j.finnegan,E.ONeill,y.yang2}@bath.ac.uk

Abstract

This work presents a novel hand pose estimation framework via intermediate dense guidance map supervision. By leveraging the advantage of predicting heat maps of hand joints in detection-based methods, we propose to use dense feature maps through intermediate supervision in a regression-based framework that is not limited to the resolution of the heat map. Our dense feature maps are delicately designed to encode the hand geometry and the spatial relation between local joint and global hand. The proposed framework significantly improves the state-of-the-art in both 2D and 3D on the recent benchmark datasets.

Keywords: hand pose estimation, dense guidance map, intermediate supervision

Introduction

In this work, we focus robust hand pose estimation from a single depth image, a challenging task due to the wide possibility of poses, missing geometric information caused by self-occlusions, and extreme viewpoints. The main idea is to facilitate deep neural nets with better knowledge of the target 3D domain, so the central task of our work is to embed geometric characteristics and spatial relationships as intermediate guidance into the framework.

We summarize our contributions as follows:

- We apply feature space supervision via dense guidance maps, which are consistent within the entire feature domain, and robust to occlusions.
- The design of our network structure combines detection based method and regression based method, and benefits from the added accuracy of intermediate predictions.
- We systematically evaluate different types of guidance maps to prove their effectiveness, achieving improved results by combining with state-of-the-art approaches.

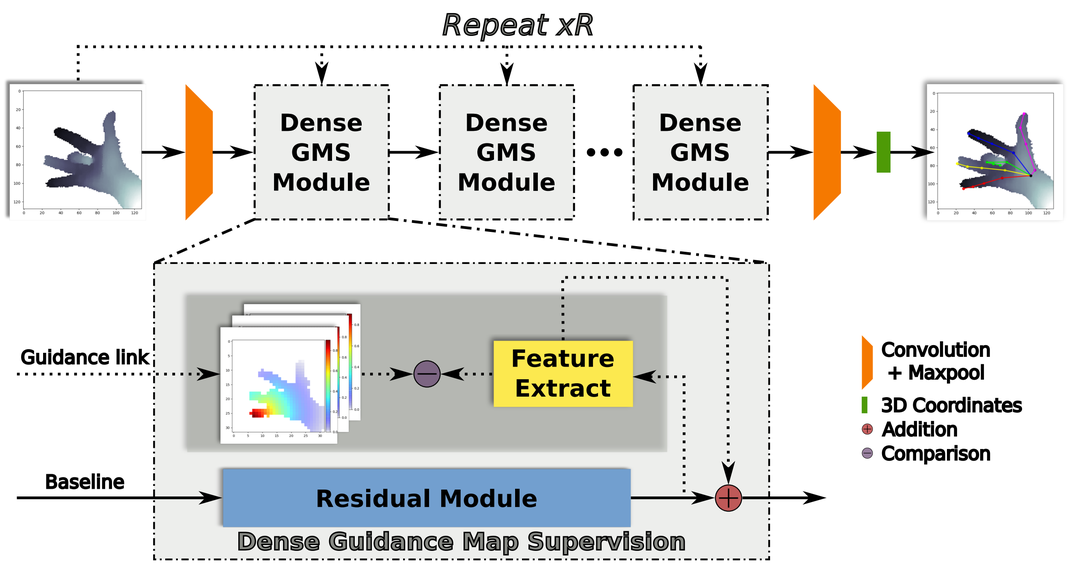

Figure 1: The pipeline of our algorithm starts from a single depth image. Our baseline method (shown in solid line) stacks R repetitions of a residual module on lower dimensional feature space, then directly regresses 3D coordinates of each joint as in a conventional CNN-based framework. In comparison, our proposed method (shown in dashed line) densely samples geometrically meaningful constraints from the input image, which provides coherent guidance to the feature representation of residual module.

Methods

Figure 1 illustrates the overall structure of our pipeline and the core Guidance Map Supervision (GMS) modules in the zoom-in. The main idea is to leverage the feature extraction effectiveness of the residual module through guidance map supervision, which further enhances the entire system's learning strength when by combining the residual link.

Guidance map supervision

The most straightforward guidance map could be easily implemented as heat-map (Figure 2 (a)), which is sparse and locally supported for each joint. But we focus more on dense guidance maps (Figure 2 (b-d)) with larger supporting neighborhood for better suppression of false positive detections.

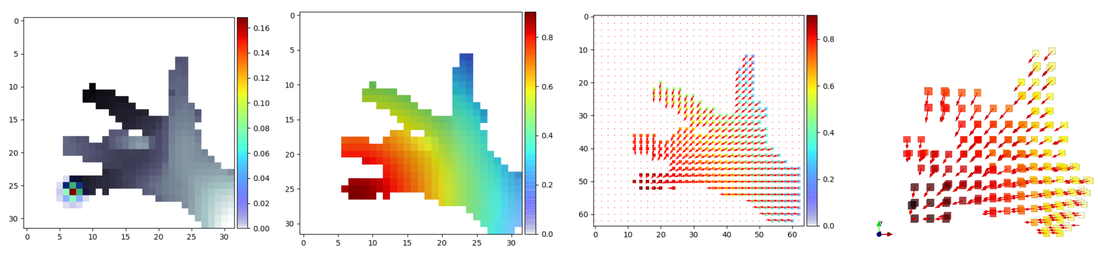

Figure 2: Different guidance maps (here we only show illustrations for the pinky fingertip). (a) 2D probability map. (b) Normalized Euclidean distance. (c-d) 2D/3D Euclidean distance plus unit offset.

We further propose geometrically meaningful dense guidance maps through approximated geodesics in 2D, as shown in Figure 3. Notice that the distance between pinky fingertip and ring fingertip is not short anymore, in contrast to using simple Euclidean distance as shown in Figure 2 (b).

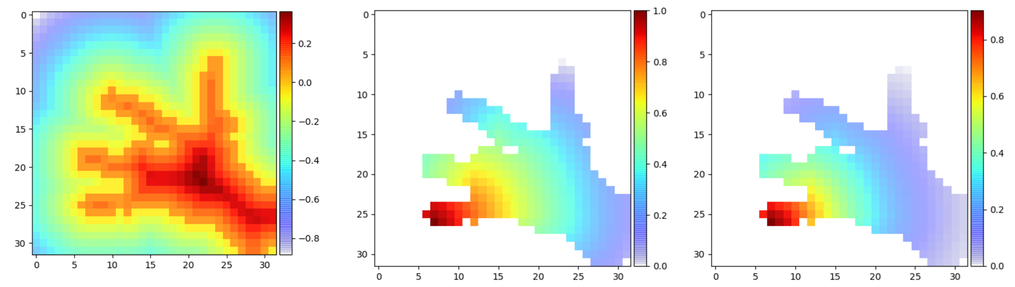

Figure 3: Geometrically more meaningful guidance maps. (a) EDT map used for propagating distance from a single point. (b-c) Two different implementations of approximate geodesic distance map for the pinky fingertip.

Results

The primary metric we use is the mean errors across all test frames for each joint, which is shown as error bars below. The second metric is the maximal per-joint error within every single frame, then the percentage curve is drawn as the ratio of correct prediction versus maximal allowed error to ground-truth annotations.

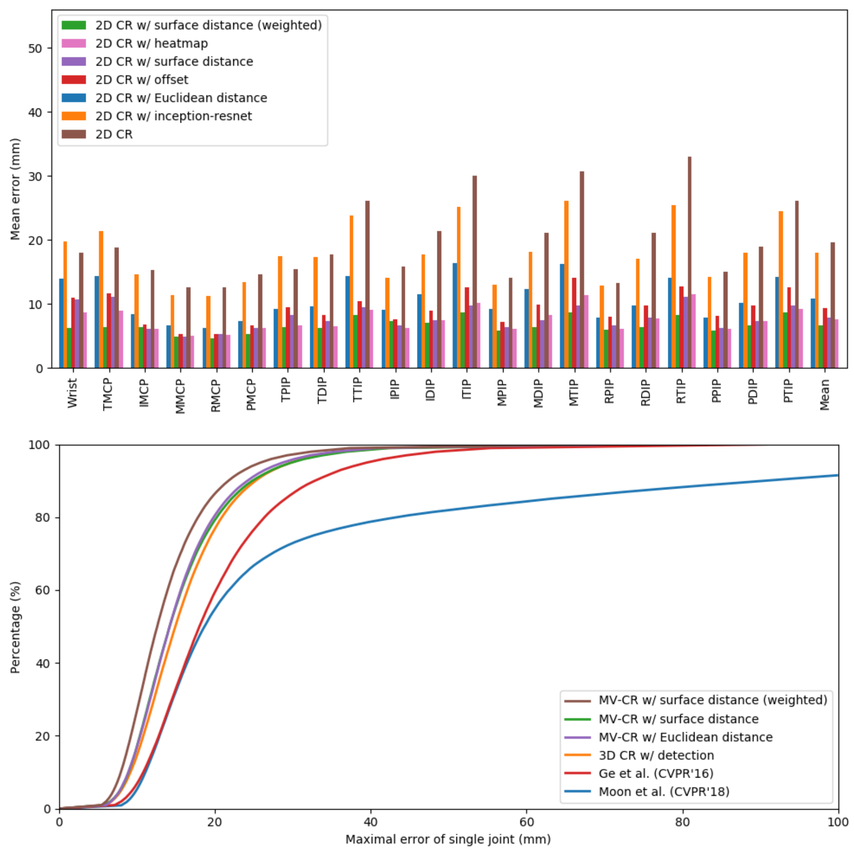

Figure 4: The upper figure shows per-joint mean errors for 2D tests, while the lower figure shows improvements over STAR methods by incorporating our core module into their frameworks. More detailed descriptions of tested methods and corresponding error statistics can be found in the table of our paper.

Conclusions

- We present a general hand pose estimation framework via intermediate supervision on dense guidance maps.
- The dense guidance maps are designed to better incorporate the geometric and spatial information of hand joints.
- We demonstrate the effectiveness of our framework and the choice of guidance maps by extensive comparisons with baseline methods in both 2D and 3D.
- Results show that our framework can robustly produce hand pose estimates, and achieve improved accuracy when combining with other STAR methods.

Forthcoming Research

Future work will explore temporal hand tracking using our framework, integrating hand detection, handle data in the wild, etc.

References

[1] Shanxin Yuan, Qi Ye, Björn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. pages 2605–2613, 2017.

Acknowledgements

We are grateful to the anonymous reviewers for their comments and suggestions. The work was supported by CAMERA, the RCUK Centre for the Analysis of Motion, Entertainment Research and Applications, EP/M023281/1.