Adopting ConvNeXt and Contextual Representation for Enhanced Feature Integration in Compact CPM for 2D Hand Pose Estimation

Sartaj Ahmed Salman †, Ali Zakir †, Hiroki Takahashi †‡
† *Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan*
‡ *Artificial Intelligence Exploration/Meta-Networking Research Center, The University of Electro-Communications, Tokyo, Japan*
s2140019@edu.cc.uec.ac.jp, a2240012@edu.cc.uec.ac.jp, and rocky@inf.uec.ac.jp

***Abstract*—The use of Convolutional Neural Networks for Hand Pose Estimation from RGB images has seen significant improvement recently. In this study, we have proposed a lightweight framework for 2D HPE, which utilizes the Convolutional Pose Machine architecture. Our approach involves using customized ConvNext as a backbone for feature extraction, along with a Global Context Block. Our experimental results demonstrate that our model outperforms existing methods on the CMU panoptic hand dataset.**

***Index Terms*—2D HPE, ConvNeXt, GCB, CPM**

I. Introduction

Human hands are the primary means of interacting with the physical world, and they play a similar role in Human-Computer Interaction. In recent years, Computer Vision (CV) has made significant advancements thanks to the growth of Artificial Intelligence and Deep Learning. Hand Pose Estimation (HPE) has also gained momentum due to these advancements, attracting the attention of researchers eager to contribute to the field. However, HPE remains challenging due to factors such as self-occlusion, dexterity, depth-ambiguity, and variations in hand size. HPE has numerous real-world applications, including Virtual and Augmented Reality, robotics, medicine, the automotive industry, and sign language processing. As 2D HPE is essential for 3D HPE, we aim to focus on 2D HPE in this paper.

We proposed a Lightweight 2D HPE model that aims to overcome computational costs while maintaining accuracy. The proposed model is designed to strike a balance between computational efficiency and accuracy.

II. Proposed Method

We utilized the ConvNeXt [1] model for feature extraction, a state-of-the-art model for many CV tasks, including HPE. Our customized ConvNeXt consists of three conv blocks, each with distinct properties. After each convolution operation, the Rectified Linear Unit activation function is applied to introduce the nonlinearity in the model. To further improve the feature extraction process, especially for the challenging task of 2D HPE, we incorporate a Global Context Block (GCB) in addition to the convolutional layers. After the feature extraction process, we applied two additional convolutional layers to adjust the number of channels in the feature maps. The resulting features are then passed to a customized CPM block to reduce the model size. Specifically, the channels in the first stage are reduced from 512 to 256, and the remaining stages are reduced from 256 to 128. This leads to a significant reduction in computational cost while maintaining high accuracy. The detailed architecture of our proposed model is shown in Fig. 1.

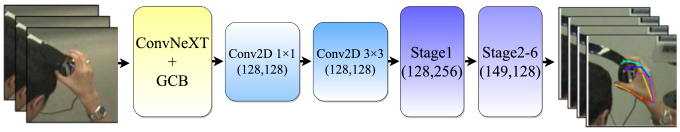

Fig. 1. Overall architecture of our proposed framework.

III. Experimental Results

Table I shows the experimental results performed on CMU panoptic hand dataset, and our proposed approach tends to perform better than the existing methods, with fewer parameters.

TABLE I
Experiment Results on CMU Panoptic Hand Dataset

Models	σ *0.04*	σ *0.08*	σ *0.12*	*Para(M)*	*GFLops*
CPM [2]	56.76	82.50	89.45	36.80	103.23
LPM-6 [3]	60.71	84.93	91.10	-	-
OCPM [4]	63.67	87.10	93.01	29.28	80.53
Our	**65.43**	**89.17**	**94.05**	**8.15**	**18.53**

*Threshold is denoted by σ

IV. Conclusion and Future Work

Our proposed framework outperforms the existing lightweight frameworks in terms of both accuracy and computational cost. We are aiming to extend our work to 3D HPE in the future.

References

[1] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.
[2] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.
[3] Y. Chen, H. Ma, D. Kong, X. Yan, J. Wu, W. Fan, and X. Xie, "Non-parametric structure regularization machine for 2D hand pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 381–390.
[4] T. Pan, Z. Wang, and Y. Fan, "Optimized convolutional pose machine for 2D hand pose estimation," *Journal of Visual Communication and Image Representation*, vol. 83, p. 103461, 2022.