Local Regression Based Hourglass Network for Hand Pose Estimation from a Single Depth Image

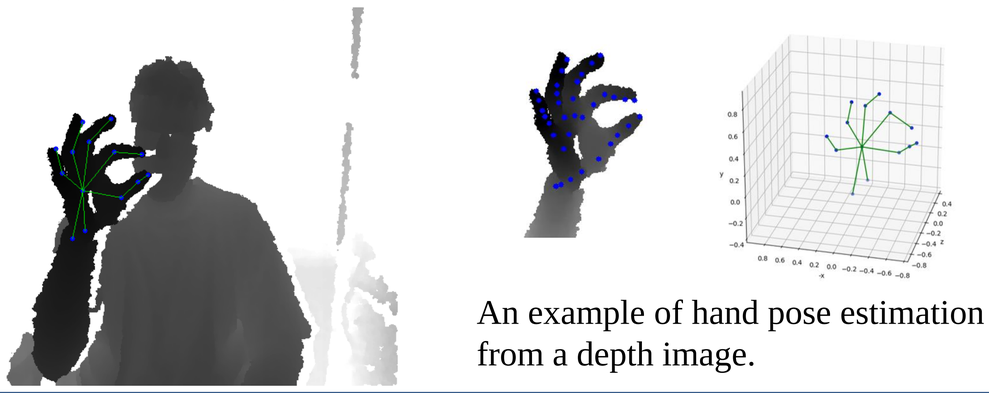
Jia Li [1] and Zengfu Wang [2]

[1] Department of Automation, University of Science and Technology of China, Hefei, China

[2] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, China

INTRODUCTION

Hand pose estimation has attracted increasing attention in recent years with the growing need in human-computer interaction, augmented reality, sign language, gesture recognition and many other potential applications. Thanks to the advent of commodity depth sensors such as Kinect, the task of accurate and real-time hand pose estimation has become more realistic. Despite much progress has been achieved, this task is still challenging due to large pose variations, many degrees of freedom, self-occlusions, device noise, etc.



An example of hand pose estimation from a depth image.

OUR WORK AND CONTRIBUTION

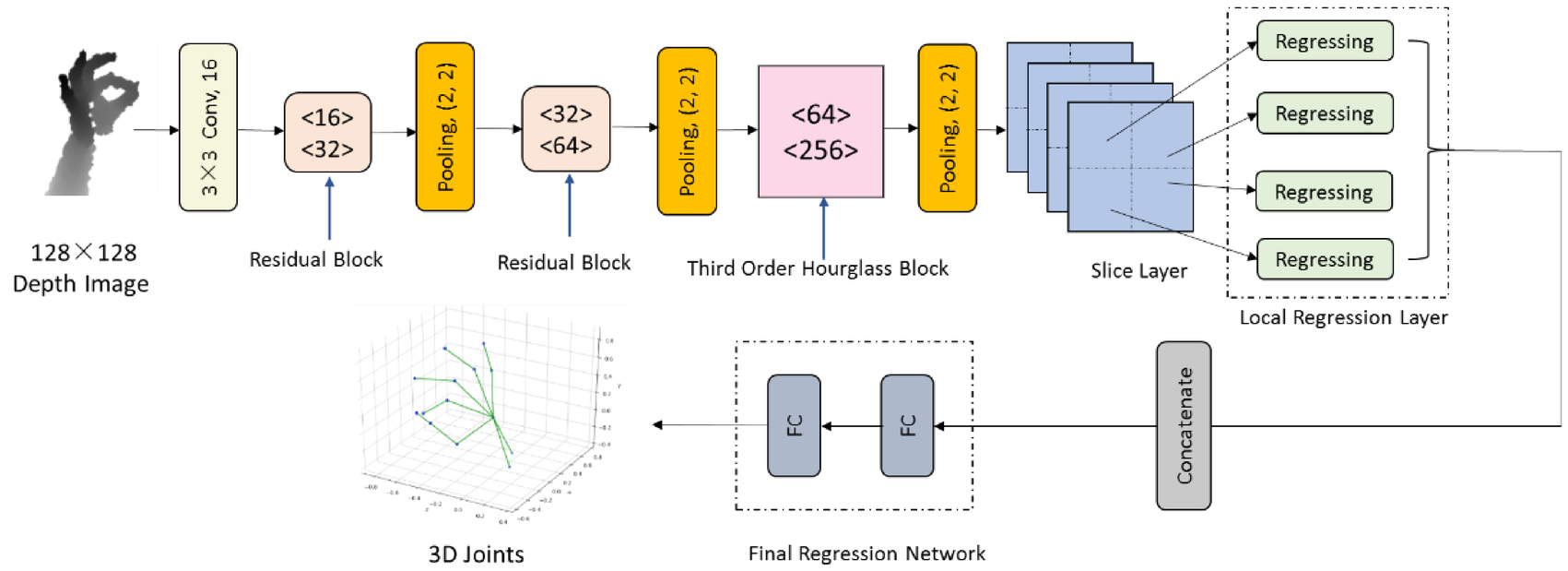
Our Work:

In this paper, we propose an end-to-end local regression based hourglass network with a modified loss function to estimate the 3D pose of the hand in a depth image. Our system can run at over 910 FPS on a single GPU, and the mean error of estimation is reduced to 12.36 mm on NYU Hand Pose Dataset (A popular hand pose dataset which is publicly available online).

Contribution:

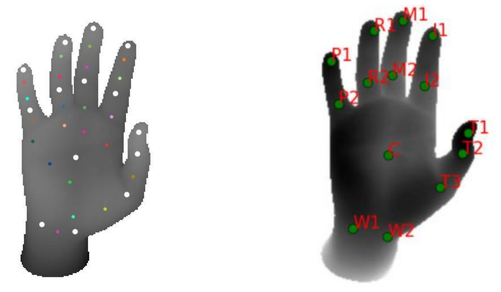
1. We build residual blocks and a third order hourglass block for feature extraction.
2. We do local regression first and then do global regression.
3. We propose a modified smooth L1 loss function which we call “max-out smooth L1 loss function”.
4. We have analyzed different loss functions and made self-comparison.

NETWORK ARCHITECTURE



Overview of our approach. We use a 3 × 3 convolutional layers (followed by a batch normalization layer and a ReLU layer), two residual blocks and a third order hourglass block for feature extraction. Subsequently, we slice the feature map into several patches and regress each patch separately. After that, we merge these regression results by concatenating them along the channel axis. The 3D coordinates of joints will be regressed by the last two fully connected layers (final regression network).

DEFINITION OF HAND JOINTS



In NYU hand pose dataset, 36 hand joints are defined, which is illustrated in the left picture. However, only 14 of them are considered during evaluation, which is shown in the right picture.

As for the preprocessing step, we followed what other work did. One joint annotation is used as the location prior of the hand in our work. The depth values and the joint coordinates are normalized to [-1, 1]. Our network produces 35 3D coordinates of joints in the normalized space.

LOSS FUNCTION

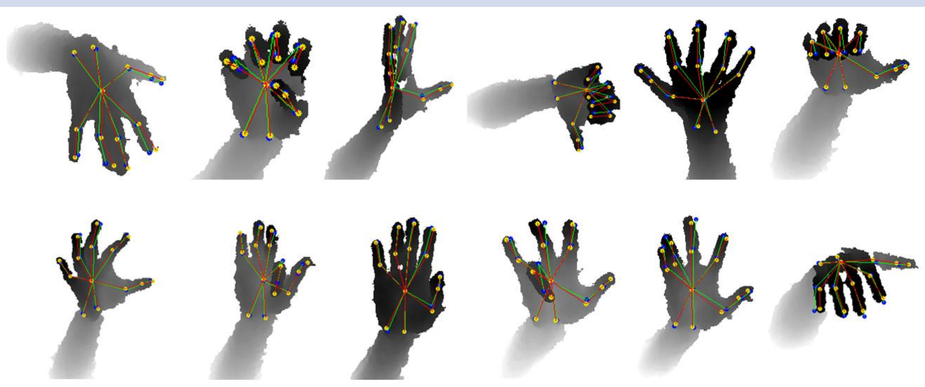
The smooth L1 norm is defined as follows:

$$|x|_{smooth} = \begin{cases} 0.5x^2, & if\ |x| \le 1 \\ |x| - 0.5, & otherwise \end{cases}$$

The Max-out Smooth L1 Loss Function is defined as:

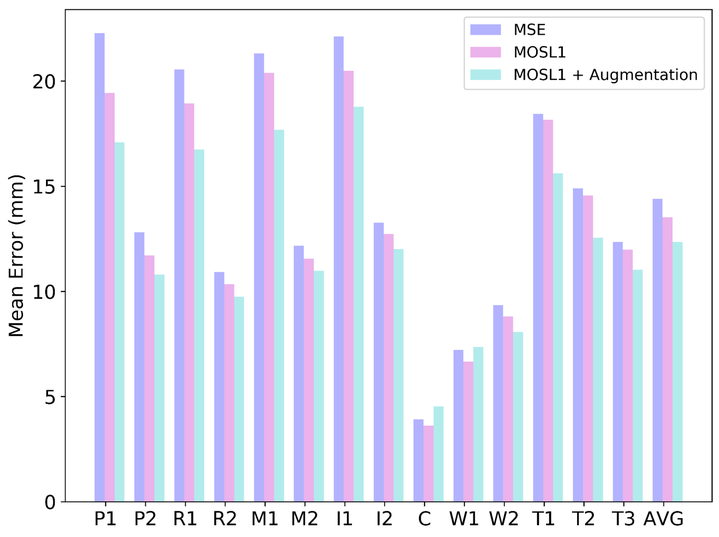
$$L_{MOSL1} = \frac{1}{N}\sum_{i=1}^{N} 3M \times \max\left(|P_i - G_i|_{smooth}\right)$$

where $\max\left(|P_i - G_i|_{smooth}\right)$ returns the biggest value of the vector $|P_i - G_i|_{smooth}$.



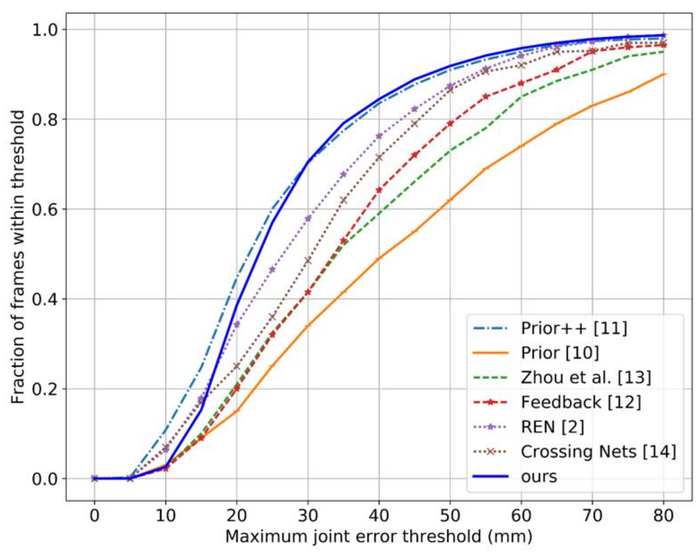
* The prediction poses are annotated with blue circles and green lines, while the ground truth poses are annotated with yellow circles and red lines.

EVALUATION



Self- comparison

The mean square loss regards all joint errors equally, while the max-out smooth L1 loss contributes to improving the accuracy of mostly error-prone joints. The mean errors descend after the max-out smooth L1 loss was used to fine tune the network.



Comparison with the State-of-the-art

The proposed approach outperforms other approaches except Prior++ according to the two evaluation metric (mean Euclidean distance error, and percentage of frames whose maximum joint error is below a threshold.). It should be noted that Prior++ only estimates 14 joints, while our network estimate 35 joints.

Paper ID: 434

Email: jialee@mail.ustc.edu.cn