3D Hand Pose Estimation from RGB image

Chih-Tien Kuo
University of Minnesota-Twin Cities
kuo00013@umn.edu

Eng Hock Lee
University of Minnesota-Twin Cities
leex7132@umn.edu

Abstract

The use of hand gesture for human-computer interaction is an attractive alternative method to the conventional interactive peripheral device such as a computer mouse. Furthermore, the advancement of the virtual environment has encouraged the development of a hand tracking system to deliver an accurate and real-time hand motion capture. Fortunately, recent development in computer vision has made it possible to capture 3D hand poses from 2D RGB images. In this proposal, we aim to improve upon the existing method of 3D hand pose estimation by introducing a biologically inspired loss function to further enhance the model generalization. We hope the outcome of our proposed method will provide a more solid and natural hand pose output compared to the previous work.

1. Introduction

Out of all human body parts, the hand is the most important input device for natural human-computer interaction (HCI) [1]. This is because it can serve as an effective and general purpose tool to allow wide-range of commands and control interfaces to the computer. The use of hand as HCI can also allow the deployment of much broader applications in sophisticated computing systems such as virtual reality [2] and augmented reality [3], making the human interaction much more intuitive and natural. More importantly, it can serve as a high degree-of-freedom control device to many demanding applications such as musical performance and surgical simulations, both which require highly sensitive and precise control to function.

A pivotal part of using the hand as a tool for HCI is the ability to accurately capture the hand motion. Currently, the most successful hand motion capturing tools are magnetic [4] or inertial [5] tracking devices. These hardware devices, also called "data glove", are able to accurately capture the global position of the hand(s) in real time. A more advanced data glove can also induce so-called haptic feedback to simulate a sense of touch to further enhance immersion experience. However, these devices tend to be bulky, expensive, and require complex calibration and setup procedure in order to obtain precise measurements.

An alternative method to capture hand motion is the computer vision based method. This method is promising because it is free of cumbersome hardware devices and can allow a more natural and non-contact interactive experience. However, accurately estimating the hand motion from images is a very challenging and computationally intensive problem, especially when dealing with the fundamental issue of self-occlusions. Fortunately, recent advancement in computer vision and deep learning techniques have made it possible to capture hand motion and estimate hand pose accurately in unprecedented speed. One such example is the work done by Ge, et al. [6], where they achieved estimating 3D hand pose and generating 3D hand shape from a single RGB image via Graph Convolutional Neural Network (CNN). In this project, we propose to improve Ge's 3D hand pose estimation by introducing a new set of loss function inspired from the biological bone structure of the human hand. We postulate that such loss function can further improve the model generalization and can result in better and more natural estimated hand pose.

2. Related Works

Many works on 3D hand pose estimation revolve around using two different types of optical sensors, namely, the RGB and the RGB-Depth sensor. While these two types of sensors only differ with additional depth level information, such information can prove useful in providing an additional layer of geometrical description for a more robust pose estimation [7, 8, 9, 10, 11, 12]. In addition, many successful implementations of hand pose estimation utilize multidimensional inputs including the use of depth and pose [13, 14].

Alternatively, information on depth can also be extracted from a single RGB image to estimate 3D hand pose. Panteleris, et al. [15] used predicted 2D hand joint location from CNN to estimate a 3D hand pose by solving the inverse kinematics problem. Similar to Panteleris, Zimmermann, et al. [16] utilized CNN to predict 2D hand joint location, but used another CNN to predict the 3D hand joint location. He,

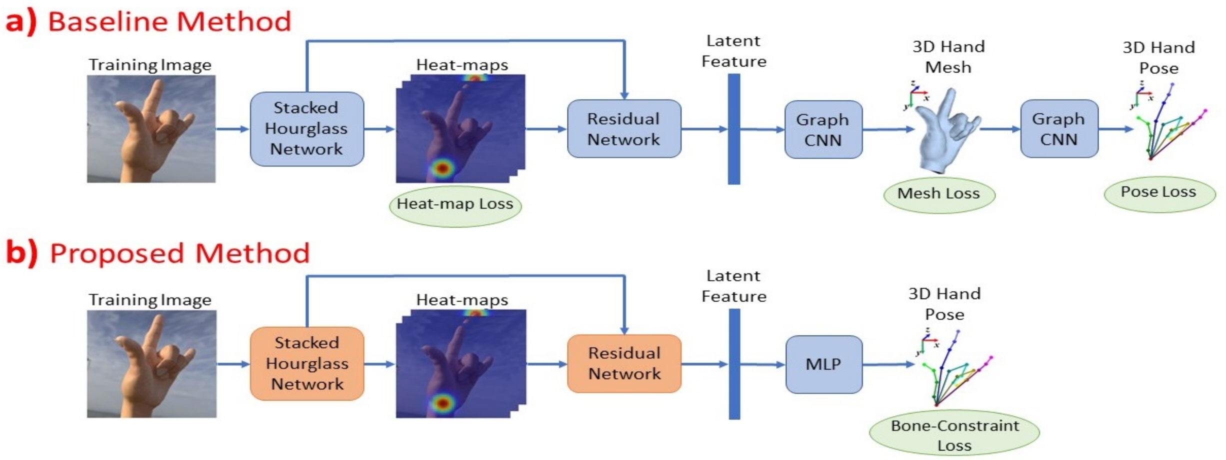

Figure 1. a) The model structure for the baseline method. b) The model structure for the proposed method. Note that the model weight of the stacked hourglass and residual networks (denoted orange box) is optimized and fixed throughout the model training.

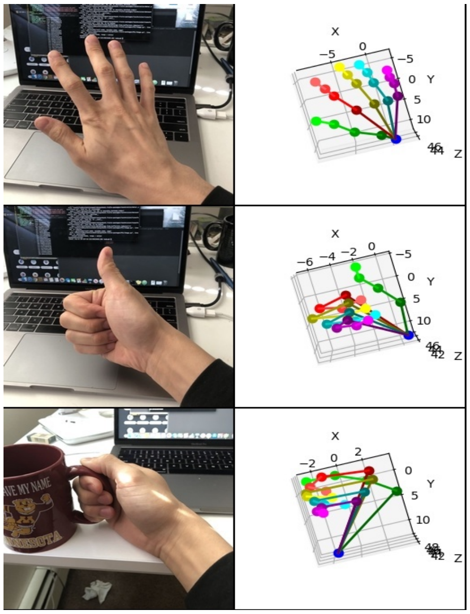

Figure 2. The original RGB image (left column) and the estimated 3D hand pose (right column) using the baseline method.

et al. [17] estimated 3D hand joint locations and then used Adverserial Learning model to fine-tune the joint locations. Recently, Ge, et al. [6], Baek, et al. [18], and Mueller, et al. [19] designed a new neural network structure that allows for the prediction of both 2D and 3D hand joint locations concurrently. The idea is to simultaneously compensate the depth ambiguities in 2D joint location by using the 3D joint information as well as resolve 3D global hand rotation from the 2D joint information. Such implementation has shown to provide excellent 3D hand pose estimation even without the use of a depth sensor.

3. Baseline Method

For the baseline method, we aim to replicate the 3D hand pose estimation method proposed in [6]. The learning model structure is described next. First, a single input RGB image is passed through a two-stacked "hourglass" network [20] to extract 2D heat-maps of 21 different hand joint positions. Then, these heat-maps are combined with the image feature maps and fitted into a residual network [21] to output an array of latent feature vector. From here, the latent feature vector is passed through a Graph CNN [22] to generate 3D hand mesh, which can then be regressed to obtain the 3D hand joint locations by using another Graph CNN. The described procedure is pictorially illustrated in Figure 1a. For our purposes, we will be using the pre-trained model provided by [6] to evaluate their result.

Figure 2 shows the qualitative results tested on the real-world image dataset (images provided by [6] for testing purposes) using the baseline method. For quantitative evaluation, we will be using the following two performance metrics: (i) pose error: the average error in Euclidean space between the estimated 3D joints and the ground truth joints; (ii) 3D PCK: the percentage of correct keypoints of which the Euclidean error distance is below a threshold. The average pose estimation for the real-world image dataset is 11.93mm. We aimed to improve upon these performance metrics obtained from the baseline method with our proposed methods.

4. Proposed Method

We propose a modification to the baseline method. This modification emphasizes in replacing the current loss function with a biologically inspired loss function. Currently, the loss function for 3D hand pose estimation used in the

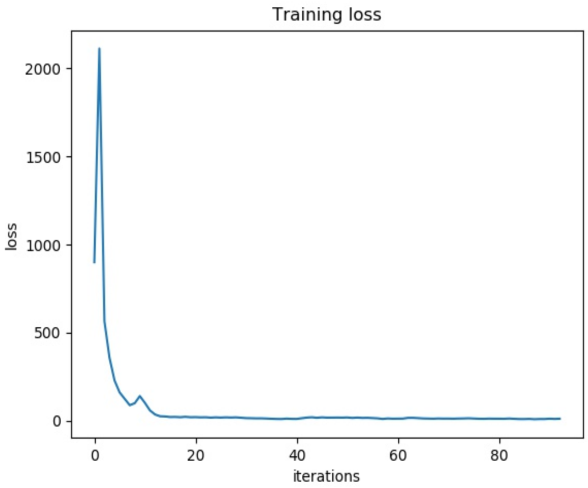

Figure 3. Training loss as a function of batch iteration.

baseline method is formulated as:

$$L_{pose} = \sum_{j=1}^{21} \left\| \boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_j \right\|_2^2 \tag{1}$$

where $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ are the ground-truth and estimated 3D joint locations respectively, and the mean squared error is summed over 21 hand joints in the 3D space. However, such loss function assumed every joints in the hand comprised of full translational degree-of-freedom, without considering the kinematics of human hand. We proposed adding additional constraints in the loss function based on bone structure similar to [17]. Specifically, the proposed bone-constrained loss function L can be formulated as:

$$L = \lambda_{pose} L_{pose} + \lambda_{len} L_{len} + \lambda_{dir} L_{dir} \tag{2}$$

where λ_{pose}, λ_{len}, and λ_{dir} are hyperparameters for the trade-off between these losses. L_{len} quantifies the distance in bone length between the ground truth and its estimates, which is defined as:

$$L_{len} = \sum_{i,j} \left| ||\boldsymbol{b}_{i,j}||_2 - ||\hat{\boldsymbol{b}}_{i,j}||_2 \right| \tag{3}$$

where $\boldsymbol{b}_{i,j} = \boldsymbol{\phi}_i - \boldsymbol{\phi}_j$ is the ground truth bone vector between joint i and j, and $\hat{\boldsymbol{b}}_{i,j}$ is its predicted counterpart. This loss function impose the translational constraint on the joints to provide a more rigid and natural hand skeleton structure. Besides, L_{dir} measures the deviation in the direction of bones:

$$L_{dir} = \sum_{i,j} \left\| \frac{\boldsymbol{b}_{i,j}}{||\boldsymbol{b}_{i,j}||_2} - \frac{\hat{\boldsymbol{b}}_{i,j}}{||\hat{\boldsymbol{b}}_{i,j}||_2} \right\| \tag{4}$$

This loss function imposes the rotational constraint on the joints such that the estimated hand pose do not look distorted. Besides modifying the loss function, we also plan to simplify the model by replacing multiple Graph CNNs to a single Multi-Layer Perceptron (MLP) network. This is because our goal focus on estimating the 3D hand pose. Therefore generating a 3D hand mesh becomes a redundant procedure. Figure 1b summarizes our proposed modeling procedure in the visual form. Note that the optimized model weights from [6] will be utilized for the hourglass and residual networks and fixed throughout the model training process, and the training procedure will be to update the MLP network weight.

5. Data Description

We will be using Stereo Tracking Benchmark Dataset (STB) [23] to train the proposed model. STB comprised of nearly 18,000 of real image taken using a stereo camera. The dataset also contain a depth dimension, but we will omit this feature in this project. Each image has a 640x480x3 resolution and contains the ground truth of 3D hand joint locations, thus making it suitable for our modeling purposes.

6. Preliminary Result

For the preliminary analysis, we want to make sure our proposed method in figure 1b can produce results that are similar to the baseline method. To do that, we begin implementing the proposed method but using only the pose loss in equation (1) as the loss function to simplify the problem. The MLP network consist of two dense layer where the first and second layer contains 512 and 256 hidden units, respectively. The output layer contains 63 units corresponds to 21 joints in 3 dimensional location. A subset of the image data from the STB dataset ($\sim$ 3000 images) is used to train the model, and the batch size is set at 32. Figure 3 show the model training loss as a function of batch iteration. The next step will be to incorporate the bone-constraint loss function into the model training.

References

[1] Erol, Ali, et al. "Vision-based hand pose estimation: A review." *Computer Vision and Image Understanding* 108.1-2 (2007): 52-73.

[2] Cameron, Charles R., et al. "Hand tracking and visualization in a virtual reality simulation." *2011 IEEE systems and information engineering design symposium*. IEEE, 2011.

[3] Malik, Shahzad, Chris McDonald, and Gerhard Roth. "Hand tracking for interactive pattern-based augmented reality." *Proceedings. International Symposium on Mixed and Augmented Reality*. IEEE, 2002.

[4] Ma, Yinghong, et al. "Magnetic hand tracking for human-computer interface." *IEEE Transactions on Magnetics* 47.5 (2011): 970-973.

[5] Gałka, Jakub, et al. "Inertial motion sensing glove for sign language gesture acquisition and recognition." *IEEE Sensors Journal* 16.16 (2016): 6310-6316.

[6] Ge, Liuhao, et al. "3d hand shape and pose estimation from a single rgb image." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2019.

[7] Yang, Xiaodong, and YingLi Tian. "Super normal vector for human activity recognition with depth cameras." *IEEE transactions on pattern analysis and machine intelligence* 39.5 (2016): 1028-1039.

[8] Oreifej, Omar, and Zicheng Liu. "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2013.

[9] Shotton, Jamie, et al. "Real-time human pose recognition in parts from single depth images." *CVPR 2011*. Ieee, 2011.

[10] Rogez, Grégory, et al. "3D hand pose detection in egocentric RGB-D images." *European Conference on Computer Vision*. Springer, Cham, 2014.

[11] Mueller, Franziska, et al. "Real-time hand tracking under occlusion from an egocentric rgb-d sensor." *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017.

[12] Garcia-Hernando, Guillermo, et al. "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[13] Dibra, Endri, et al. "Monocular RGB hand pose inference from unsupervised refinable nets." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

[14] Baek, Seungryul, et al. "Kinematic-layout-aware random forests for depth-based action recognition." *arXiv preprint arXiv:1607.06972* (2016).

[15] Panteleris, Paschalis, Iason Oikonomidis, and Antonis Argyros. "Using a single RGB frame for real time 3D hand pose estimation in the wild." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

[16] Zimmermann, Christian, and Thomas Brox. "Learning to estimate 3d hand pose from single rgb images." *Proceedings of the IEEE International Conference on Computer Vision.* 2017.

[17] He, Yiming, et al. "3D Hand Pose Estimation in the Wild via Graph Refinement under Adversarial Learning." *arXiv preprint arXiv:1912.01875* (2020).

[18] Baek, Seungryul, Kwang In Kim, and Tae-Kyun Kim. "Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[19] Mueller, Franziska, et al. "Ganerated hands for real-time 3d hand tracking from monocular rgb." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[20] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *European conference on computer vision*. Springer, Cham, 2016.

[21] He, Kaiming, et al. "Deep residual learning for image recognition." [**?**]. 2016.

[22] Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst. "Convolutional neural networks on graphs with fast localized spectral filtering." *Advances in neural information processing systems*. 2016.

[23] Zhang, Jiawei, et al. "A hand pose tracking benchmark from stereo matching." *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.