

Point-to-Point Regression PointNet for 3D Hand Pose Estimation

Liuhaio Ge¹, Zhou Ren², and Junsong Yuan³

¹ Institute for Media Innovation, Interdisciplinary Graduate School,
Nanyang Technological University, Singapore
ge0001ao@e.ntu.edu.sg

² Snap Inc., 64 Market Street, Venice, CA, USA
zhou.ren@snapchat.com

³ Department of Computer Science and Engineering,
State University of New York at Buffalo, NY, USA
jsyuan@buffalo.edu

Abstract. Convolutional Neural Networks (CNNs)-based methods for 3D hand pose estimation with depth cameras usually take 2D depth images as input and directly regress holistic 3D hand pose. Different from these methods, our proposed Point-to-Point Regression PointNet directly takes the 3D point cloud as input and outputs point-wise estimations, i.e., heat-maps and unit vector fields on the point cloud, representing the closeness and direction from every point in the point cloud to the hand joint. The point-wise estimations are used to infer 3D joint locations with weighted fusion. To better capture 3D spatial information in the point cloud, we apply a stacked network architecture for PointNet with intermediate supervision, which is trained end-to-end. Experiments show that our method can achieve outstanding results when compared with state-of-the-art methods on three challenging hand pose datasets.

Keywords: 3D Hand Pose Estimation

1 Introduction

A key technology for human-computer interaction in virtual reality and augmented reality applications is accurate and real-time 3D hand pose estimation, which allows direct hand interaction with virtual objects. Despite the recent progress of 3D hand pose estimation with depth cameras [23, 13, 51, 38, 43, 36, 35, 11, 22, 45, 54, 17], it remains challenging to achieve accurate and robust results due to the high dimensionality and large variations of 3D hand pose, high similarity among fingers, severe self-occlusion, and noisy depth images.

Most of the recently proposed 3D hand pose estimation methods [11, 22, 45, 10, 53, 12, 19, 4, 5] are based on convolutional neural networks (CNNs) and have achieved drastic performance improvement on large hand pose datasets [43, 36, 35, 55]. Many methods directly regress 3D coordinates of hand joints or hand pose parameters using CNNs [7, 11, 9, 22, 45, 12, 19, 4, 5, 21, 56]. However, the direct mapping from input representation to 3D hand pose is highly non-linear and

