International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01856-0

Efficient Annotation and Learning for 3D Hand Pose Estimation: A Survey

Takehiko Ohkawa[1] · **Ryosuke Furuta**[1] · **Yoichi Sato**[1]

Received: 9 June 2022 / Accepted: 12 July 2023

Abstract
In this survey, we present a systematic review of 3D hand pose estimation from the perspective of efficient annotation and learning. 3D hand pose estimation has been an important research area owing to its potential to enable various applications, such as video understanding, AR/VR, and robotics. However, the performance of models is tied to the quality and quantity of annotated 3D hand poses. Under the status quo, acquiring such annotated 3D hand poses is challenging, e.g., due to the difficulty of 3D annotation and the presence of occlusion. To reveal this problem, we review the pros and cons of existing annotation methods classified as manual, synthetic-model-based, hand-sensor-based, and computational approaches. Additionally, we examine methods for learning 3D hand poses when annotated data are scarce, including self-supervised pretraining, semi-supervised learning, and domain adaptation. Based on the study of efficient annotation and learning, we further discuss limitations and possible future directions in this field.

Keywords Hand pose estimation · Efficient annotation · Learning with limited labels

1 Introduction

The acquisition of 3D hand pose annotations[1] has presented a significant challenge in the study of 3D hand pose estimation. This makes it difficult to construct large training datasets and develop models for various target applications, such as hand-object interaction analysis (Boukhayma et al., 2019; Hampali et al., 2020), pose-based action recognition (Iqbal et al., 2017; Tekin et al., 2019; Sener et al., 2022), augmented and virtual reality (Liang et al., 2015; Han et al., 2022; Wu et al., 2020), and robot learning from human demonstration (Ciocarlie & Allen, 2009; Handa et al., 2020; Qin et al., 2022; Mandikal & Grauman, 2021). In these application scenarios, we must consider methods for annotating hand data, and select an appropriate learning method according to the amount and quality of the annotations. However, there is currently no established methodology that can give annotations efficiently and learn even from imperfect annotations. This motivates us to review methods for building training datasets and developing models in the presence of these challenges in the annotation process.

During the annotations, we encounter several obstacles including the difficulty of 3D measurement, occlusion, and dataset bias. As for the first obstacle, annotating 3D points from a single RGB image is an ill-posed problem. While annotation methods using hand markers, depth sensors, or multi-view cameras can provide 3D positional labels, these setups require a controlled environment, which limits available scenarios. As for the second obstacle, occlusion hinders annotators from accurately localizing the positions of hand joints. As for the third obstacle, annotated data are biased to a specific condition constrained by the annotation method. For instance, annotation methods based on hand markers or multi-view setups are usually installed in laboratory settings, resulting in a bias toward a limited variety of backgrounds and interacting objects.

[1] We denote 3D pose as the 3D keypoint coordinates of hand joints, $\mathrm{P}^{3D} \in \mathbb{R}^{J\times 3}$ where J is the number of joints.

Communicated by Dima Damen.

✉ Takehiko Ohkawa
ohkawa-t@iis.u-tokyo.ac.jp

Ryosuke Furuta
furuta@iis.u-tokyo.ac.jp

Yoichi Sato
ysato@iis.u-tokyo.ac.jp

[1] Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

Published online: 07 August 2023