DOI: 10.3217/978-3-85125-524-9-18

Feedback Loop and Accurate Training Data for 3D Hand Pose Estimation[†]

Markus Oberweger[1], Gernot Riegler[1], Paul Wohlhart[1] and Vincent Lepetit[1,2]

Abstract— **In this work, we present an entirely data-driven approach to estimating the 3D pose of a hand given a depth image. We show that we can correct the mistakes made by a Convolutional Neural Network (CNN) trained to predict an estimate of the 3D pose by using a feedback loop of Deep Networks, also utilizing a CNN architecture.**

Since this approach critically relies on a training set of labeled frames, we further present a method for creating the required training data. We propose a semi-automated method for efficiently and accurately labeling each frame of a depth video of a hand with the 3D locations of the joints.

I. INTRODUCTION

Accurate hand pose estimation is an important requirement for many Human Computer Interaction or Augmented Reality tasks. Due to the emergence of 3D sensors, there has been an increased research interest in hand pose estimation in the past few years [3], [6], [7]. Despite the additionally available information from 3D sensors, it is still a very challenging problem, because of the large number of degrees of freedom, and because images of hands exhibit self-similarity and self-occlusions.

A popular approach to predict the position of the joints is to use a discriminative method [3], [7], which are now robust and fast. To further refine the pose, such methods are often used to initialize a complex optimization where a 3D model of the hand is fit to the input depth data [5].

In this paper, we build upon recent work that learns to generate images from training data [1] in order to remove the requirement of a 3D hand model. We introduce a method that learns to provide updates for improving the current estimate of the pose, given the input image and the image generated for this pose estimate. Running these steps iteratively, we can correct the mistakes of an initial estimate provided by a simple discriminative method.

However, this approach, amongst other recent work (*e.g.* [6], [7]), has shown that a large amount of accurate training data is required for reliable and precise pose estimation. Although having accurate training data is very important, there was only limited scientific interest in the creation of such, and authors have had to rely on *ad hoc* ways that are prone to errors [6]. These errors result in noisy training and test data, and make training and evaluating uncertain. Therefore, we developed a semi-automated approach that makes it easy to annotate sequences of articulated poses in 3D from a single depth sensor only.

In the next two sections, we first describe our proposed feedback loop, and then we present our method for efficiently creating training data.

II. TRAINING A FEEDBACK LOOP

We aim at estimating the pose of a hand in the form of the 3D locations of its joints from a single depth image. We assume that a training set of depth images labeled with the corresponding 3D joint locations is available. An overview of our method is shown in Fig. 1.

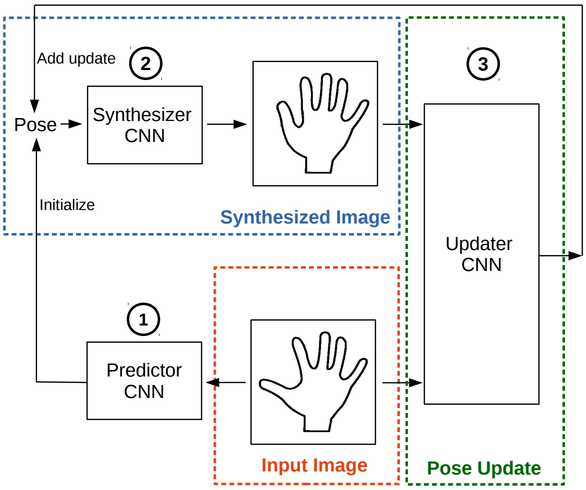

Fig. 1. Overview of our method: We use a CNN (1) to predict an initial estimate of the 3D pose given an input depth image of the hand. The pose is then used to synthesize an image (2), which is used together with the input depth image to derive a pose update (3). The update is applied to the pose and the process is iterated.

We first train a *predictor* to predict an initial pose estimate in a discriminative manner given an input depth image. We use a Convolutional Neural Network to implement this predictor with a very simple architecture [3].

In practice, the initial pose is never perfect, and following the motivation provided in the introduction, we introduce a hand model learned from the training data. This CNN-based model, referred to as *synthesizer*, can synthesize the depth image corresponding to a given pose. The network architecture is strongly inspired by [1]. It predicts an initial latent representation of feature maps, followed by subsequent unpooling and convolution layers to generate a depth image.

Further, we introduce a third function that we call the *updater*. It learns to predict updates to improve the pose estimate, given the input image and the image produced

[†] This work is based on published work in ICCV'15 [4] and CVPR'16 [2].

[1] The authors are with the Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria {oberweger,riegler,wohlhart,lepetit}@icg.tugraz.at

[2] The author is with the Laboratoire Bordelais de Recherche en Informatique, Université de Bordeaux, Bordeaux, France

by the synthesizer. We iterate this update several times to improve the initial pose estimate. Again, the updater function is implemented as a CNN. The architecture is inspired by a Siamese network with two identical paths. One path is fed with the observed depth image and the second path is fed with the image from the synthesizer.

Ideally, the output of the updater should bring the pose estimate to the correct pose in a single step, which is a very difficult problem in practice. However, our only requirement from the updater is to predict an update resulting in a pose closer to the ground truth. The introduction of the synthesizer allows us to virtually augment the training data and to add arbitrary poses to the set of poses, which the updater might perceive during testing and be asked to correct. We refer to our paper [4] for more details.

III. Creating Training Data Efficiently

Since the presented hand pose estimation method critically relies on labeled training frames, we present a method for the creation of such frames. Given a sequence of depth maps capturing a hand in motion, we want to estimate the 3D joint locations for each depth map with minimal effort.

We start by automatically selecting a subset of depth frames, we will refer to as *reference frames*, for which a user is asked to provide annotations. Our method selects these reference frames based on the appearances of the frames over the whole sequence. For this, we train an autoencoder that learns an unsupervised representation that is sensitive to image nuances due to hand articulation. We use this representation to formalize the frame selection as a submodular optimization. A user is then asked to provide the 2D reprojections of the joints with visibility information in these reference frames, and whether these joints are closer or farther from the camera than the parent joint in the hand skeleton tree. We use this information to automatically recover the 3D locations of the joints by solving a least-squares problem. Next, we iteratively propagate these 3D locations from the reference frames to the remaining frames. We initialize the pose of the frame with the pose of the visually closest reference frame and optimize the local appearance together with spatial constraints. This gives us an initialization for the joint locations in all the frames. However, each frame is processed independently. We can improve the estimates further by introducing temporal constraints on the 3D locations and perform a global optimization, enforcing appearance, temporal, and spatial constraints over all 3D locations for all frames. If this inference fails for some frames, the user can still provide additional 2D reprojections; by running the global inference again, a single additional annotation typically fixes many frames. See our paper [2] for more details.

IV. Evaluation

We evaluate our hand pose estimation method on the NYU Hand Pose Dataset [7], a challenging real-world benchmark for hand pose estimation. This dataset is publicly available, is backed up by a huge quantity of annotated samples, and also shows a high variability of poses.

We show the benefit of using our proposed feedback loop to increase the accuracy of the 3D joint localization in Fig. 2. While [7] and [3] have an average 3D joint error of 21 mm and 20 mm respectively, our proposed method reaches an error reduction to 16.5 mm. The initialization with the simple and efficient proposed predictor has an error of 27 mm. When we use a more complex initialization [3] with an error of 23 mm, we can decrease the average error to 16 mm.

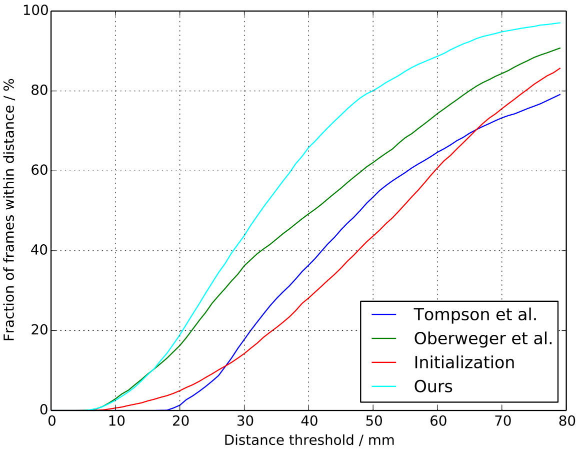

Fig. 2. Quantitative evaluation of hand pose estimation. The figure shows the fraction of frames where all joints are within a maximum distance. A higher area under the curve denotes better results. We compare our method to the baseline of Tompson *et al.* [7] and Oberweger *et al.* [3]. Although our initialization is worse than both baselines, we can boost the accuracy of the joint locations using our proposed feedback loop.

To demonstrate our training data creation approach, we evaluate it on a synthetic dataset, which is the only way to have depth maps with ground truth 3D locations of the joints. On this dataset we evaluate the accuracy of the automatically inferred 3D locations for the reference frames. We obtain an average 3D joint error of 3.6 mm only from 2D reprojections with visibility and depth order. Our method is also robust to annotation noise. We can propagate the 3D joint locations to the remaining frames, for which we achieve an average 3D joint error of 5.5 mm over the full sequence by only requiring manual 2D annotations for 10% of all frames.

References

[1] A. Dosovitskiy, J. T. Springenberg, and T. Brox, "Learning to Generate Chairs with Convolutional Neural Networks," in *CVPR*, 2015.

[2] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit, "Efficiently Creating 3D Training Data for Fine Hand Pose Estimation," in *CVPR*, 2016.

[3] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands Deep in Deep Learning for Hand Pose Estimation," in *Proc. of CVWW*, 2015.

[4] ——, "Training a Feedback Loop for Hand Pose Estimation," in *ICCV*, 2015.

[5] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and Robust Hand Tracking from Depth," in *CVPR*, 2014.

[6] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim, "Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture," in *CVPR*, 2014.

[7] J. Tompson, M. Stein, Y. LeCun, and K. Perlin, "Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks," *ACM Transactions on Graphics*, vol. 33, no. 5, pp. 169–179, 2014.