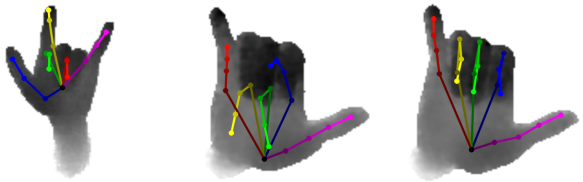
Efficiently Creating 3D Training Data for Fine Hand Pose Estimation

Markus Oberweger Gernot Riegler Paul Wohlhart Vincent Lepetit
Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{oberweger,riegler,wohlhart,lepetit}@icg.tugraz.at

Introduction Recent work on articulated pose estimation [1, 7, 8, 9] has shown that a large amount of accurate training data makes reliable and precise estimation possible. For human bodies, Motion Capture [1] can be used to generate large datasets with sufficient accuracy. However, creating accurate annotations for hand pose estimation is far more difficult, and still an unsolved problem. Motion Capture is not an option anymore, as it is not possible to use fiducials to track the joints of a hand. Moreover, the human hand has more degrees of freedom than are generally considered for 3D body tracking, and an even larger amount of training data is probably required.

The appearance of depth sensors has made 3D hand pose estimation easier, but has not solved the problem of the creation of training data entirely. Despite its importance, the creation of a training set has been overlooked so far, and authors have had to rely on *ad hoc* ways that are prone to errors, as shown in Fig. 1. Complex multi-camera setups [6, 9] together with tracking algorithms have typically been used to create annotations. For example, Tompson *et al*. [9] used a complex camera setup with three RGBD cameras to fit a predefined 3D hand model. Looking closely at the resulting data, it seems that the 3D model was often manually adjusted to fit the sequences better and in between these manually adjusted frames the fit can be poor. Further, the dataset of [8] contains many misplaced annotations, as discussed by [3]. Although recent datasets [7] have paid more attention to high quality annotations, they still contain annotation errors. These errors result in noisy training and test data, and make training and evaluating uncertain.



(a) (b) (c)

Figure 1: Recent hand pose datasets exhibit significant errors in the 3D locations of the joints. (a) is from the ICVL dataset [8], and (b) from the MSRA dataset [7]. Both datasets were annotated by fitting a 3D hand model, which is prone to converge to a local minimum. In contrast, (c) shows the annotations acquired with our proposed method for the same frame as in (b). (Best viewed in color)

Creating Training Data Efficiently For all of these reasons, we developed a semi-automated approach that makes it easy to annotate sequences of articulated poses in 3D. 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. Fig. 2 shows an overview of our approach. We start by automatically selecting some of the depth maps we will refer to as *reference frames*. 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 [4], which we refer to as *z-order*. This can be done easily and quickly, and 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 annotator can still provide additional 2D reprojections; by running the

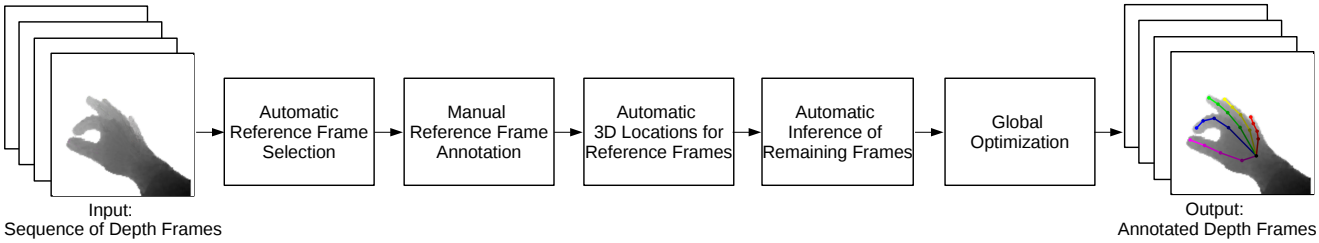


Figure 2: Overview of our method. We start by automatically selecting a subset of frames that have to be manually annotated. Then we automatically infer the 3D locations from the annotations. Given the 3D locations for the reference frames, we propagate these locations to the remaining frames, and run a global optimization over the full sequence. (Best viewed in color)

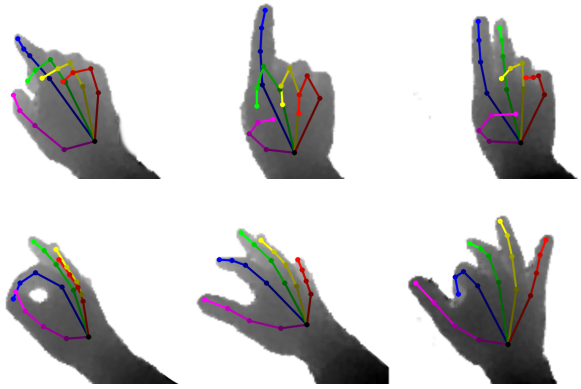


Figure 3: Egocentric 3D hand pose estimation is an appealing feature for different Augmented Reality or Human Computer Interaction applications. Our method made it possible to create a fully annotated dataset of more than 2000 frames from an egocentric viewpoint, which is considered to be very challenging [5]. (Best viewed in color)

global inference again, a single additional annotation typically fixes many frames. We refer to our paper [2] for more details. We will make the full code and dataset available on our website.

Evaluation We evaluate our approach using both synthetic data and real images. We first 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 show, that our proposed reference frame selection can be used to efficiently select the frames that maximize pose coverage and simultaneously minimize the number of frames, *i.e*. annotation work. Further, we evaluate the accuracy of the automatically inferred 3D locations for the reference frames. We obtain an average Euclidean 3D joint error of 3.6 mm only from 2D reprojections with visibility and z-order. Our method is also robust to annotation noise. We then show, that we can propagate the 3D joint locations to the remaining frames. We achieve an average 3D joint error of 5.5 mm over the full sequence by only requiring manual annotations for 10% of all frames. We evaluate the impact of the number of reference frames, and even for 1% annotated frames the average 3D error is only 7.2 mm.

We then provide a qualitative evaluation on real images. We show that we can improve the annotations of existing datasets, which yield more accurate predicted poses. We use the recent MSRA dataset [7], where we show that better annotations improve the accuracy of a state-of-the-art 3D hand pose estimation method [3]. As Fig. 3 shows, our approach also allows us to provide the first fully annotated egocentric sequences, with more than 2000 frames in total.

Conclusion Given the recent developments in Deep Learning, the creation of training data may now be the main bottleneck in practical applications of Machine Learning for hand pose estimation. Our method brings a much needed solution to the creation of accurate 3D annotations of hand poses. It avoids the need for motion capture systems, which are cumbersome and cannot always be used. Moreover, it could also be applied to any other articulated structures, such as human bodies.

References

[1] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 2014.

[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] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for Monocular Human Pose Estimation. In *CVPR*, 2014.

[5] G. Rogez, M. Khademi, J. S. Supancic, J. Montiel, and D. Ramanan. 3D Hand Pose Detection in Egocentric RGB-D Images. In *ECCV*, 2014.

[6] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive Markerless Articulated Hand Motion Tracking Using RGB and Depth Data. In *ICCV*, 2013.

[7] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded Hand Pose Regression. In *CVPR*, 2015.

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

[9] J. Tompson, M. Stein, Y. LeCun, and K. Perlin. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. *ACM Transactions on Graphics*, 33, 2014.