Hybrid One-Shot 3D Hand Pose Estimation by Exploiting Uncertainties

Georg Poier[1], Konstantinos Roditakis[2,3], Samuel Schulter[1]
{poier,schulter}@icg.tugraz.at/croditak@ics.forth.gr
Damien Michel[2], Horst Bischof[1], Antonis A. Argyros[2,3]
{michel,argyros}@ics.forth.gr/bischof@icg.tugraz.at

[1] Institute for Computer Graphics and Vision
Graz University of Technology, Austria

[2] Institute of Computer Science
FORTH, Greece

[3] Computer Science Department
University of Crete, Greece

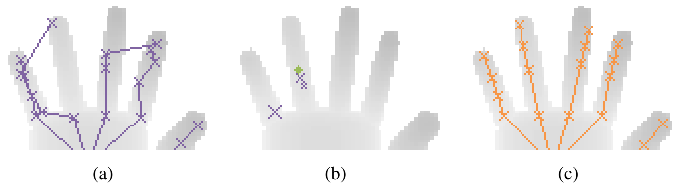

Figure 1: (a) A learned joint regressor might fail to recover the pose of a hand due to ambiguities or lack of training data. (b) We make use of the inherent uncertainty of a regressor by enforcing it to generate multiple proposals. The crosses show the top three proposals for the proximal interphalangeal joint of the ring finger for which the corresponding ground truth position is drawn in green. The marker size of the proposals corresponds to degree of confidence. (c) A subsequent model-based optimisation procedure exploits these proposals to estimate the true pose.

Traditionally, the task of hand pose estimation was approached mostly by model-based or data-driven schemes. Model-based approaches have been shown to perform well in a wide range of scenarios. However, they require initialisation and cannot recover easily from tracking failures that occur due to fast hand motions. Data-driven approaches, on the other hand, can quickly deliver a solution, but the results often suffer from lower accuracy or missing anatomical validity compared to those obtained from model-based approaches. We propose to combine the merits of both schemes in a hybrid approach. This way, the method provides anatomically valid and accurate solutions without requiring manual initialisation or suffering from track losses.

Main Idea For the task of 3D hand pose estimation, the substantial similarities between individual fingers and complex finger interactions cause ambiguities and uncertainties which are often disregarded by previous works. In contrast, we have the model-based step exploiting the inherent uncertainties of the data-driven part. First, a learned regressor is employed to deliver multiple initial hypotheses for the 3D position of each hand joint. These proposals approximate the distribution of joint positions under the learned model and thus capture the uncertainty of the model. Subsequently, the parameters of an anatomically valid hand pose are found by model-based optimisation which exploits the uncertainties captured by the proposal distributions. To do this, the optimisation is privy to internal information from the learned regressor. In this way failures of the regressor can be corrected during optimisation (see Fig. 1).

Proposal Generation For the generation of an approximated proposal distribution we build upon a prominent approach for body pose estimation [2], which has also been previously adapted for hand pose estimation [5]. The approach relies on Random Forests [1] to infer a 3D distribution of likely hand joint locations. Using the discriminative Random Forest based method, inference of the individual joint proposals is completely independent from the other joints. While, in this way, the complex dependencies do not need to be modeled, the resulting proposals are not necessarily compatible with anatomical constraints.

Optimisation In order to obtain a valid pose we employ a predefined model of a hand. Such a model can be specified by a number of parameters defining the global position and orientation of the hand as well as the joint angles for each finger joint. During optimisation these parameters are constrained based on anatomical studies [3] which avoids impossible configurations. The goal of the optimisation is to find the model parameters which best describe the modes of the proposal distributions, obtained from the regression forest, while still obeying anatomic constraints.

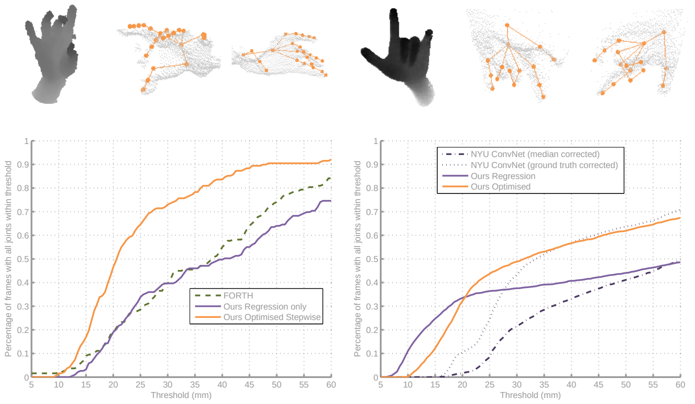

Figure 2: Top row: Qualitative results for two input depth maps rendered under different viewpoints. Bottom row: The ratio of frames with *all* joints within a certain distance to the ground truth as a function of this distance. Comparing the proposed approach (orange) to regression only (purple) and [4] (green, dotted) in the left figure on a synthetic sequence, and to the CNN based hybrid approach [6] (dotted, dark purple) in the right figure on the NYU dataset. See the paper for more results.

Results Experimental results on various datasets (ICVL, NYU, synthetic) show that our specific combination of a data-driven and model-based scheme outperforms state-of-the-art representatives of the model-based, data-driven and hybrid paradigms (see, *e.g.* Fig. 2). In line with this, we found it especially effective to make the optimisation scheme aware of the uncertainty of the regressor. Moreover, using a simple observation we were able to make the optimisation converge much faster: Given the global orientation of the hand, the fingers can move almost independently of each other. This allows us to split the optimisation problem into sub-problems of lower dimensionality, which resulted in a dramatic speed-up without loss of accuracy. See the paper and our website[1] for more details and more results.

Acknowledgment This work has been supported by the EU projects 3D-PITOTI (ICT-2011-600545), FP7-IP-288533 Robohow and FP7-ICT-2011-9 WEARHAP, and by the Austrian Research Promotion Agency under the FIT-IT Bridge programme, project TOFUSION (838513).

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Ross B. Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew W. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Proc. ICCV*, 2011.

[3] John Y. Lin, Ying Wu, and Thomas S. Huang. Modeling the constraints of human hand motion. In *Proc. Workshop on Human Motion*, 2000.

[4] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *Proc. BMVC*, 2011.

[5] Danhang Tang, Tsz-Ho Yu, and Tae-Kyun Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *Proc. ICCV*, 2013.

[6] Jonathan Tompson, Murphy Stein, Yann LeCun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. on Graphics*, 33(5):169:1–169:10, 2014.

[1] Online resources at `http://lrs.icg.tugraz.at/research/hybridhape/`