Cascaded Hand Pose Regression

Xiao Sun[1,2], Yichen Wei[1], Shuang Liang[3], Xiaoou Tang[2], Jian Sun[1]
[1]Microsoft Research. [2]Chinese University of Hong Kong. [3]Tongji University.

Abstract We extends the previous 2D cascaded object pose regression work [3] in two aspects so that it works better for 3D articulated object. Our first contribution is 3D pose-indexed features that generalize the previous 2D parameterized features and achieve better invariance to 3D transformations. Our second contribution is a principled hierarchical regression that is adapted to the articulated object structure. It is therefore more accurate and faster. Comprehensive experiments verify the state-of-the-art accuracy and efficiency of the proposed approach on the challenging 3D hand pose estimation problem, on a public dataset and our new dataset.

Introduction The problem of pose estimation of 3D articulated objects such as human body and hand has been studied for decades. Recent years have seen rapid progress and significant success of human body pose estimation [6] using consumer depth sensors. The state-of-the-art learning approach [6] classifies depth pixels into body parts and then infer the body pose from the pixel classification result. This paradigm has been applied for hand pose estimation [2, 4, 7, 13] but is less successful than for body pose. This is because body is mostly near-frontal and there is less occlusion between limbs. However, hand motion exhibits much larger variations in both camera viewpoints and finger articulations. This generates more complex depth images and makes the pixel classification much more difficult. Furthermore, the pixel classification approaches do not capture the structural constraints in the hand pose.

Regression based approaches directly estimate the hand pose from the depth image, using latent regression forest [11] or deep convolutional neutral networks [12]. Such methods are more principled since their learning is directly guided by the task. Nevertheless, only one regression model is learnt in such works, which may have insufficient capacity to model the complex image variations, especially under large viewpoints and hand motions.

We present a cascaded regression approach that is more robust under large viewpoints and complex hand poses. It is directly motivated by the cascaded pose regression framework [3], where the object pose is estimated progressively via a sequence of weak regressors and each weak regressor uses features that depend on the estimated pose from the previous stage. Such *pose indexed features* provide better geometric invariance and simplify the learning tasks. This framework has been successfully applied to several 2D pose estimation tasks [1, 3]. Yet, it is unclear how to use it for 3D objects with complex articulated structure like human hand.

We extend the framework for 3D articulated object and propose novel techniques to address new issues absent before. Our first contribution is *3D pose-indexed features*. While we use the similar pixel difference features as in cascaded pose regression [1, 3] and many previous works [2, 4, 6, 7, 13], we show that the parameterization of pixel indexing is the key to achieve certain geometrical invariance and analyze the invariance properties of previous features [6, 13]. We explain our rationale about a new 3D parameterization, which generalizes the previous methods and achieves better invariance to 3D transformations.

Our second contribution is a principled *hierarchical* approach that is adapted for the structure of articulated objects. Our key observation is that different object parts typically exhibit different amount of variations and degrees of freedom due to the articulated structure. Thus, regressing all parts together is unnecessarily difficult and causes slow convergence and degraded accuracy. Our hierarchical approach *regresses the pose of different parts sequentially in the order of their articulation complexity*. It firstly estimates the pose of the easier root part (such as palm). Estimation of more difficult sub-parts (such as fingers) are then conditioned on the pose of the root part and thus easier. The hierarchical approach does not only converge faster but is also more accurate.

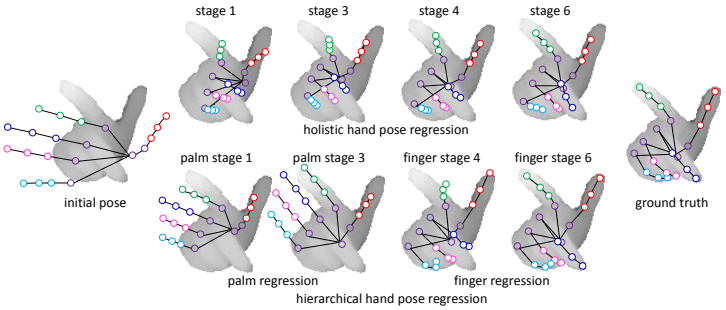

Figure 1: Illustration of cascaded hand pose regression on a real example. Starting from the depth image and a rough initial hand pose, the hand pose is iteratively updated through six stages and approaching the ground truth. The top row is the holistic hand regression. The bottom row is the hierarchical regression, that is, palm is updated in the first three stages with fingers fixed (relatively to the palm) and fingers are updated in the last three stages with palm fixed.

The proposed approach works on general 3D articulated objects. It is applied for hand pose estimation in this work, as exemplified in Fig. 1. Comprehensive experiments show that it significantly outperforms the state-of-the-art, on both public data in [11] and a large challenging dataset collected by us. In addition to the high accuracy, our regression is also very fast ($>$ 300 FPS on CPU, single thread). It is thus highly complementary to model based hand tracking approaches [5, 8, 9, 10] (to initialize tracking) and would be influential for real applications.

[1] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.

[2] C.Keskin, F.Kirac, Y.E.Kara, and L.Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV*, 2012.

[3] Piotr Dollar, Peter Welinder, and Piotro Perona. Cascaded pose regression. In *CVPR*, 2010.

[4] D.Tang, T.Y, and T.K.Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *ICCV*, 2013.

[5] I.Oikonomidis, N.Kyriazis, and A.A.Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, 2011.

[6] J.Shotton, A.Fitzgibbon, M.Cook, T.Sharp, M.Finocchio, R.Moore, A.Kipman, and A.Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[7] Hui Liang, Junsong Yuan, and Daniel Thalmann. Parsing the hand in depth images. *IEEE Trans. Multimedia*, 2014.

[8] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *CVPR*, 2014.

[9] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew Fitzgibbon, and Shahram Izadi. Accurate, robust, and flexible realtime hand tracking. In *CHI*, 2015.

[10] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *ICCV*, 2013.

[11] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *CVPR*, 2014.

[12] Jonathan Tompson, Murphy Stein, Yann LeCun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. Graph.*, 2014.

[13] Chi Xu and Li Cheng. Efficient hand pose estimation from a single depth image. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

Our dataset is available at http://research.microsoft.com/en-us/people/yichenw