Model-based Deep Hand Pose Estimation

Xingyi Zhou[1], Qingfu Wan[1], Wei Zhang[1], Xiangyang Xue[1], Yichen Wei[2]

[1]Fudan University, [2]Microsoft Research

[1]{zhouxy13, qfwan13, weizh, xyxue}@fudan.edu.cn, [2]yichenw@microsoft.com

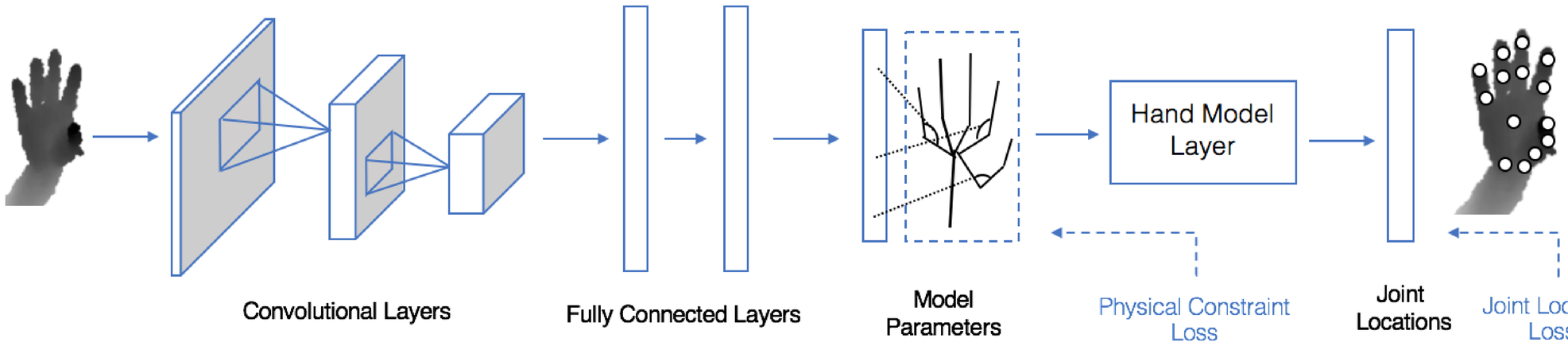


Figure 1: Illustration of model based deep hand pose learning. After standard convolutional layers and fully connected layers, the hand model pose parameters (mostly joint angles) are produced. A new hand model layer maps the pose parameters to the hand joint locations via a forward kinematic process. The joint location loss and a physical constraint based loss guide the end-to-end learning of the network.

Goal

Given a depth image of human hand, estimate accurate 3D joint locations.

Challenges

- Highly articulated structure
- Significant self-occlusion
- various viewpoint changes

Previous Approaches

Model based(Generative)

- Synthesize observation from hand geometry.
- Optimize the discrepancy to obtain the pose.
- Accurate but slow.

Learning based(Discriminative)

- Learn a direct regression function that maps the image appearance to hand pose.
- Efficient but suffer from invalid poses.

Hybrid Discriminative and Generative

- Discriminative method for initialization.
- Model based refinement.
- Separated multi-stages.

Our Approach

We propose a model based deep learning approach that fully exploits the hand model geometry. We develop a new layer that realizes the non-linear forward kinematics, that is, mapping from the joint angles to joint locations. The layer is efficient, differentiable, parameter-free and servers as an intermediate representation in the network.

Contribution

- For the first time, we show that the end-to-end learning using the non-linear forward kinematics layer in a deep neutral network is feasible for hand pose estimation.
- we show that using joint location loss and adding an additional regularization loss on the intermediate pose representation are important for accuracy and pose validity.

Code is available at

https://github.com/tenstep/DeepModel

Method

Hand Model

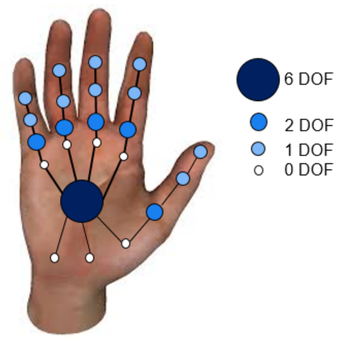
A hand model is a map from hand pose parameters Θ to 3D joint locations Y

$\mathcal{F} : \mathcal{R}^D \rightarrow \mathcal{R}^{J\times 3}$

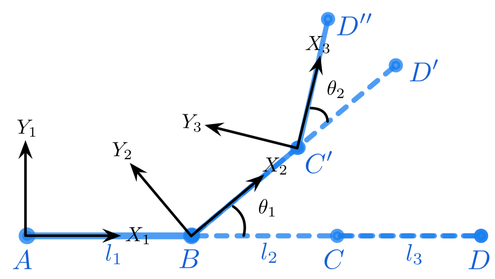
$D = 26$: The degrees of freedom of human hand

$J = 23$: The number of key joints.

$Y = \mathcal{F}(\Theta)$



Forward Kinematic



$\mathbf{p}_{u^{(k)}} = (\prod_{t\in Pa(u)} Rot_{\phi_t}(\theta_t) \times Trans_{\phi_t}(\theta_t))[0,0,0,1]^\top$

Deep Learning with a Hand Model Layer

Joint location loss:

$$L_{jt}(\Theta) = \frac{1}{2}||\mathcal{F}(\Theta) - Y||^2$$

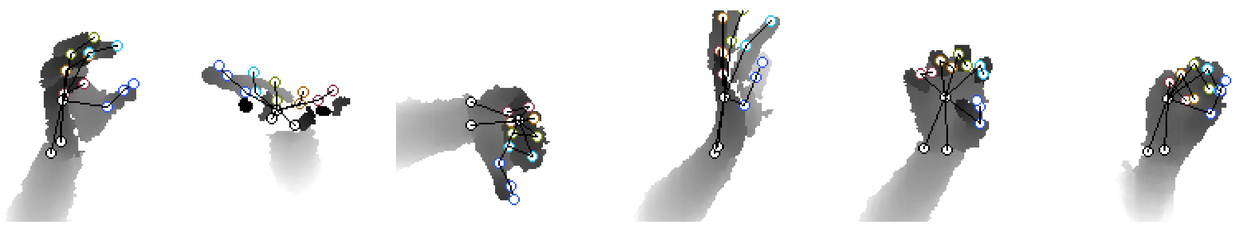
Physical constraint loss:

$$L_{phy}(\Theta) = \sum_i [max(\underline{\theta_i} - \theta_i, 0) + max(\theta_i - \overline{\theta_i}, 0)].$$

Overall loss:

$$L(\Theta) = L_{jt}(\Theta) + \lambda L_{phy}(\Theta)$$

Experiments



Self Comparison

Methods \ Metrics	Joint error	Angle error
direct joint	17.2mm	21.4°
direct parameter	26.7mm	12.2°
ours w/o phy	**16.9mm**	**12.0°**
ours	**16.9mm**	12.2°

We use a PSO based off-line model fitting to obtain joint angle ground truth.

- Direct joint is hard to be fitted in a model.
- Direct parameter has large joint error.
- Ours w/o phy is the best, but there are 18.6% frames have out-of-range angles.
- Physical constraint ensures pose validity.

Comparison with the State-of-the-art

NYU Dataset [4, 1, 2]

ICVL Dataset[3, 1]

References

[1] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. *CVWW*, 2015.

[2] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a feedback loop for hand pose estimation. In *ICCV*, 2015.

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

[4] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 2014.