Model-based Deep Hand Pose Estimation

Xingyi Zhou, Qingfu Wan, Wei Zhang,
Xiangyang Xue, Yichen Wei

Fudan University & Microsoft Research

July 7, 2016

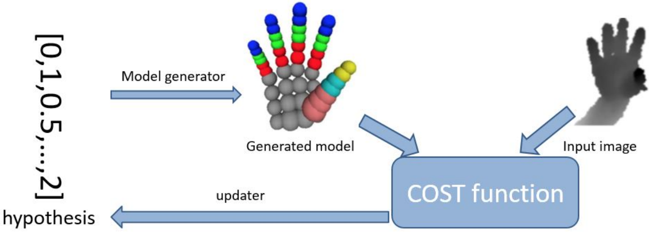
Motivation

- Various applications in human-computer interaction, augmented reality and driving analysis ...
- Widely used commercial depth sensors.
- Hot research topic.

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

Generative Approaches

Model-based, synthesize and optimize.

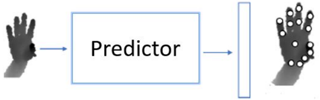


- [Oikonomidis et al., 2011]
- [Makris et al., 2015]
- [Qian et al., 2014]
- [Tagliasacchi et al., 2015]
- [Sharp et al., 2015]

- Could be highly accurate
- Guaranteed to be valid
- Slow

Discriminative Approaches

Learning-based, learn a direct regression function.



Random Forest Regressor

- [Keskin et al., 2012]
- [Tang et al., 2013]
- [Xu and Cheng, 2013]
- [Sun et al., 2015]
- [Li et al., 2015]

CNN Regressor

- [Oberweger et al., 2015a]

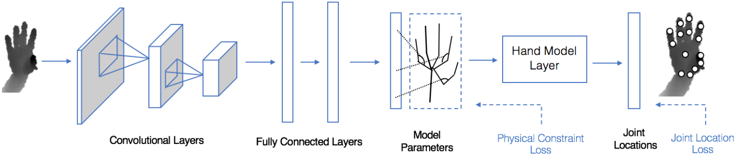
- Much more efficient
- Results are coarse
- Violate hand geometry

Hybrid Approaches

Use discriminative method for initialization, and model-based refinement.

- [Tompson et al., 2014]
- [Oberweger et al., 2015b]
- [Dong et al., 2015]
- [Sridhar et al., 2015]

Model-based Deep Hand Pose Estimation

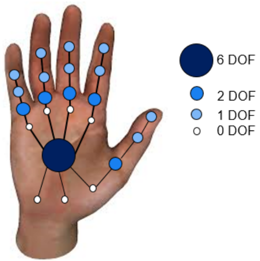


- We designed a novel layer in deep learning that realized the non-linear forward kinematic mapping from joint angles to joint locations.
- We add a physical constraint as a multi-task loss in the objective function to ensure physical validity.

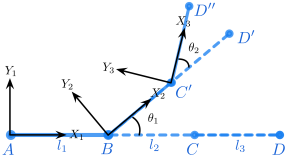
Hand Model

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

- $\mathcal{F} : \mathcal{R}^D \to \mathcal{R}^{J \times 3}$
- $D = 26$: The DOF of human hand
- $J = 23$: The number of key joints
- $Y = \mathcal{F}(\Theta)$
- $\theta_i \in [\underline{\theta_i}, \overline{\theta_i}]$



Forward Kinematics

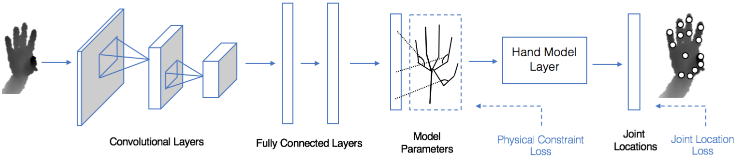


$$\mathbf{p}_{u^{(k)}} = \Big(\prod_{t \in Pa(u)} Rot_{\phi_t}(\theta_t) \times Trans_{\phi_t}(\theta_t) \Big) [0,0,0,1]^\top$$

$$p_{u^{(k)}} = \left(\prod_{t \in Pa(u)} \begin{bmatrix} \cos(\theta_z^t) & -\sin(\theta_z^t) & 0 & 0 \\ \sin(\theta_z^t) & \cos(\theta_z^t) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta_x^t) & -\sin(\theta_x^t) & 0 \\ 0 & \sin(\theta_x^t) & \cos(\theta_x^t) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\theta_y^t) & 0 & \sin(\theta_y^t) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta_y^t) & 0 & \cos(\theta_y^t) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & p_x^t \\ 0 & 1 & 0 & p_y^t \\ 0 & 0 & 1 & p_z^t \\ 0 & 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Global Z Rotation X Rotation Y Rotation Translation Local

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:

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

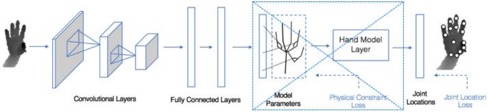
Self-Comparison

NYU Hand Pose Dataset:

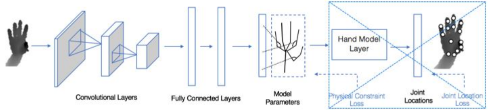
- Accurate joint locations annotation.
- We use an off-line model fitting to obtain angles ground truth.

Baselines:

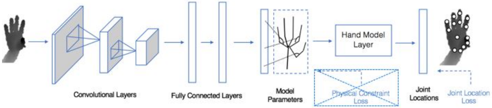
- direct joint regression



- direct parameter regression

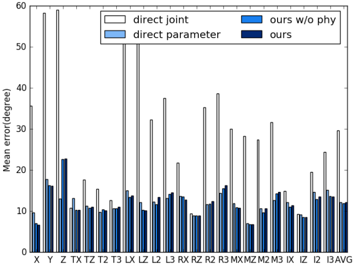


- without physical constraint



Self-Comparison(Results)

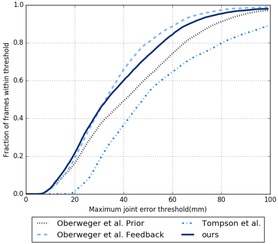
Methods \ Metrics	Joint error	Angle error
direct joint	17.2*mm*	21.4°
direct parameter	26.7*mm*	12.2°
ours w/o phy	**16.9mm**	$\mathbf{12.0^{\circ}}$
ours	**16.9mm**	12.2°

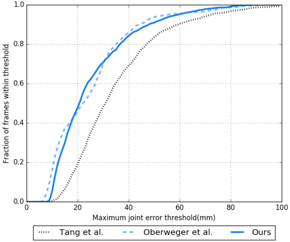


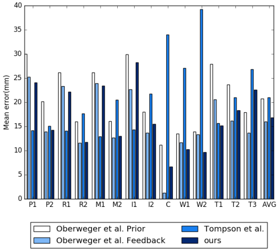
Results:

- 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 reduces invalid frames to 0.9%.

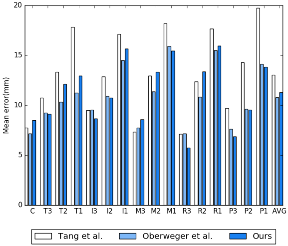
Comparison with the State-of-the-art







NYU Dataset



ICVL Dataset

Conclusion

- End-to-end learning using the non-linear forward kinematics layer in a deep neutral network is feasible for hand pose estimation.
- Adding an additional regularization loss on the intermediate pose representation is important for pose validity.
- Exploit the prior knowledge in learning process.

Q & A

Code is available at
https://github.com/tenstep/DeepModel

{zhouxy13, qfwan13, weizh, xyxue}@fudan.edu.cn
yichenw@microsoft.com