Model-based Hand Pose Estimation for Generalized Hand Shape with Spatial Transformer Network

Shile Li*[1], Jan Wöhlke*[1] and Dongheui Lee[1,2]

[1]Technical University of Munich [2]German Aerospace Center, *co-first authors

1 Introduction

Since the emergence of large annotated datasets [1], state-of-the-art hand pose estimation methods have been mostly based on discriminative learning [2]. Recently, a hybrid approach has embedded a kinematic layer into the deep learning structure in such a way that the pose estimates obey the physical constraints of human hand kinematics [3]. However, the existing approach relies on a single person's hand shape parameters, which are fixed constants. Therefore, the existing hybrid method has problems to generalize to new, unseen hands. In this work, we extend the kinematic layer to make the hand shape parameters adaptable. In this way, the learnt network can generalize towards arbitrary hand shapes. Furthermore, we show that by applying Spatial Transformer Network [4], the performance of a regression task can be also improved. The effectiveness and limitations of our proposed approach are evaluated on the Hands 2017 challenge dataset [1].

2 Method

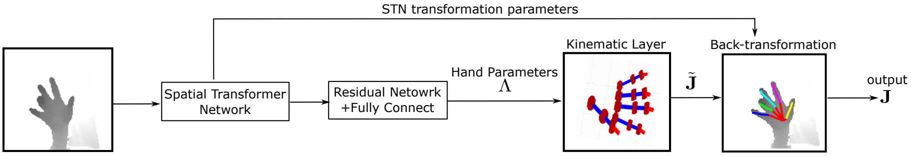
STN transformation parameters
Spatial Transformer Network
Residual Netowrk +Fully Connect
Hand Parameters Λ
Kinematic Layer
$\tilde{\mathbf{J}}$
Back-transformation
output $\mathbf{J}$

Fig. 1. Overview of our approach. The input images are appearance-normalized using Spatial Transformer Network. Then, hand parameters $\mathbf{\Lambda}$ are estimated, which are fed into a kinematic layer that maps $\mathbf{\Lambda}$ to joint locations. Finally, the joint locations are back-transformed into the initial coordinate system.

Kinematic Layer with variable hand shape parameter

The kinematic hand model layer implements the forward kinematics of the hand and therefore represents a mapping from hand parameters $\mathbf{\Lambda}$ to 3D joint locations $\tilde{\mathbf{J}}$. It is parameter free. In combination with the a residual network, the whole network can be trained end-to-end.

The inputs $\mathbf{\Lambda}$ to the kinematic layer consists of four groups: 6D global pose, which we define as the hand base $\mathbf{b}$, Wrist and finger base positions in the hand coordinate system $\mathbf{v}_{Wrist}$, $\{\mathbf{v}_i\}_{i=_1^5}$, 15 finger bone lengths $\{r_{i,n}\}_{i=_1^5,n=_1^3}$, 25 finger joint angles $\{\theta_{i,n}\}_{i=_1^5,n=_1^5}$.

The 3D joint locations $\tilde{\mathbf{J}}$ are calculated using the hand parameters $\mathbf{\Lambda}$ by chaining the appropriate transformation matrices. For example, the 3D joint location of the each finger tip's position $\tilde{\mathbf{j}}_{i,TIP}$ is

$$\tilde{\mathbf{j}}_{i,TIP} = \mathbf{T}_{\text{BASE}}\left(\mathbf{b}\right)\mathbf{T}_{\text{VEC},i}\left(\mathbf{v}_i\right)\prod_{n=1}^{5}\mathbf{T}_{\text{DH},n}\left(\theta_i, r_i\right)\left(0\ 0\ 0\ 1\right)^T, \tag{1}$$

where the kinematics of the fingers are modeled using the DH convention with joint angle limits.

Embedding the kinematic layer into the whole system (Fig. 1), the total loss term for training is:

$$L_{total} = L_{joints} + L_{fingerBase} + L_{boneLength} + L_{angleConstr} \tag{2}$$

where L_{joints}, $L_{fingerBase}$ and $L_{boneLength}$ are Euclidean distance between the estimated results and ground truth and $L_{angleConstr}$ is a penalization term for joint angles if they violates physical valid limits.

Spatial Transformer Network for hand pose regression

We also apply a Spatial Transformer Network (STN) [4] for the input image. The STN estimates a 2x3 matrix $\mathbf{T}_{STN} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ to transform the image to a similar distribution of appearance. In order to preserve the validity of hand kinematic layer, the spatial transformation parameters are constraint to rotation, scaling and translation, where $a_{11} = a_{22}$ and $a_{12} = -a_{21}$. This is achieved by estimating the scaling factor s, sine and cosine value of the rotation angle α separately: $\mathbf{T}_{STN} = s\begin{bmatrix} cos_\alpha & -sin_\alpha & a_{13} \\ sin_\alpha & cos_\alpha & a_{23} \end{bmatrix}$, where we force $cos_\alpha^2 + sin_\alpha^2 = 1$. During training, with the estimated STN parameters, the corresponding ground truth data are also transformed accordingly for correct loss computation. For inference, the estimated joint positions can be also back-transformed in a straight forward way.

3 Results

Our method is evaluated on the Hands 2017 Challenge dataset [1]. It is currently the largest and most diverse dataset available. Its training set contains 957032 depth images of five different hands. Therefore, it is suited for learning to regress hand parameters for various hand shapes. The test set consists of 295510 depth images of ten different hand shapes, of which five are the same as in the training set and five are entirely new.

Table 1 compares the pose estimation result of different methods. The Kinematic version achieves good accuracy of 12.96 mm, indicating that our kinematic layer can generalize to different hand shapes successfully. Adding the STN, the

Table 1. Average per joint error of different models

Approach	Avg test [mm]	Seen test [mm]	Unseen test[mm]
Direct Regression	10.97	8.98	12.62
Kin	12.98	10.71	14.86
STN+Kin	12.12	9.93	13.95

error reduces to 12.12 mm, showing that the STN can be also applied to improve regression task. The STN shows an appearance normalizing effect on our hand pose estimation task (Fig. 2), where the STN tends to rotate the hand such that the hand points to the upper-right direction. The direct regression of hand pose using Residual Net achieves even lower error of 10.97 mm. However, if we fit the result of direct regression using inverse kinematic, we found out that there are 8.32% of joint angles that violate physical limit of human hand (Table 2). Using our approach with kinematic constraints, the joint violation is reduced to 0.057%.

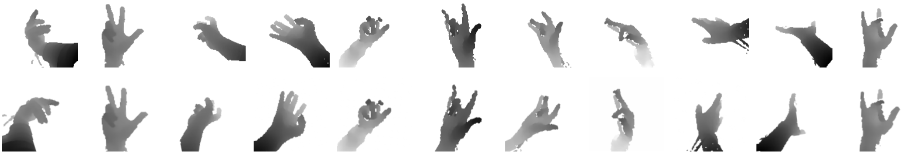

Fig. 2. Appearance normalizing effect of STN: First row: input images. Second row: transformed image after STN

Table 2. Joint angle constraint violation with and without kinematic hand model layer

Approach	Violated joint limits	Avg violation in case of violation	Avg violation in total
Direct Regression	8.32%	35.57°	5.77°
Our Approach	0.057%	0.51°	0.00°

References

1. Yuan, S., Ye, Q., Garcia-Hernando, G., Kim, T.K.: The 2017 hands in the million challenge on 3d hand pose estimation. arXiv:1707.02237 (2017)
2. Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J.Y., Lee, K.M., Molchanov, P., Kautz, J., Honari, S., Ge, L., et al.: 3d hand pose estimation: From current achievements to future goals. arXiv:1712.03917 (2017)
3. Zhou, X., Wan, Q., Zhang, W., Xue, X., Wei, Y.: Model-based deep hand pose estimation. In: Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press (2016) 2421–2427
4. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. (2015) 2017–2025