Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation

Qi Ye*, Shanxin Yuan*, Tae-Kyun Kim

Department of Electrical and Electronic Engineering,
Imperial College London
{q.ye14,s.yuan14,tk.kim}@imperial.ac.uk

1 Structures of the CNN Models

We present all the CNN models in our experiment section. Fig 1 and Fig 2 illustrate the CNN models used in our method. Fig 3 and Fig 4 are the CNN models for the baseline Holi and Holi_Derot. Fig 5 shows the overall structure of the baseline Holi_SA and Fig 5 demonstrates the CNN models used in all the parts of Holi_SA.

In Fig 1 and Fig 2, the features λ_{lj}^k for the CNN $f_{lj}^k (k > 0)$ is transformed from the feature maps in the initial CNN and the transformation results are max-pooled with pool size 4×4, 2×2, 2×2 for different resolution channels. Thus, the size of output feature maps of the spatial attention module before maxpooling is 32×32, 16×16, 8×8.

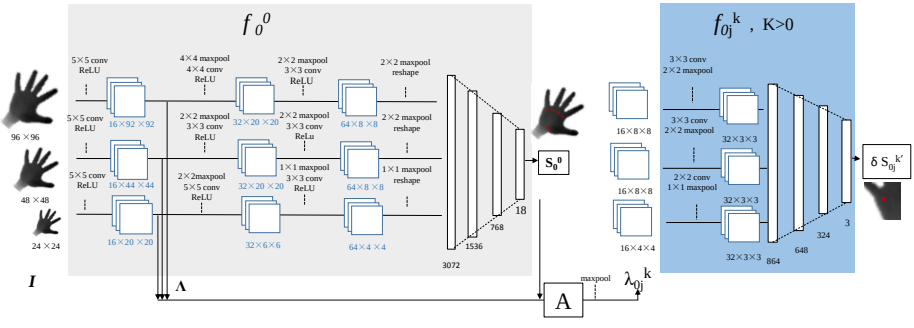

Fig. 1. Structure of the CNN models in Layer 0 of our method. Left: the CNN regressing all locations of the joints in Layer 0 in the initial stage. Right: the CNN for a single joint in the refinement stage k. The features for the CNN are transformed from the feature maps in the initial CNN and the transformation results are maxpooled with pool size 4×4, 2×2, 2×2 for different resolution channels.

* indicates equal contribution

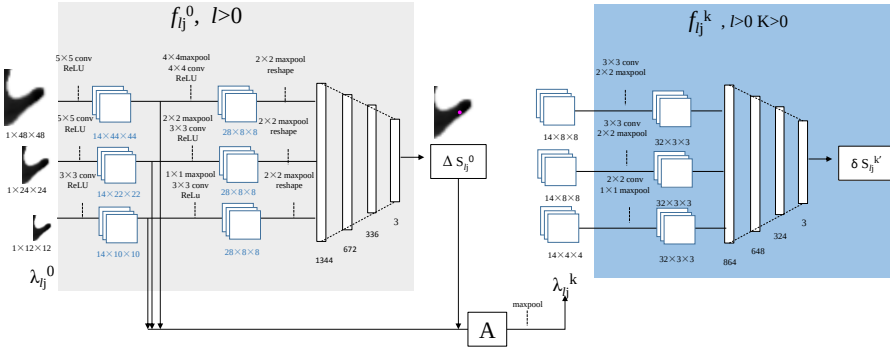

Fig. 2. Structure of the CNN models in Layer $l(l = 1, 2, 3)$ of our method. Left: the CNN regressing a single joint in the initial stage. Right: a CNN for a single joint in the refinement stage k.

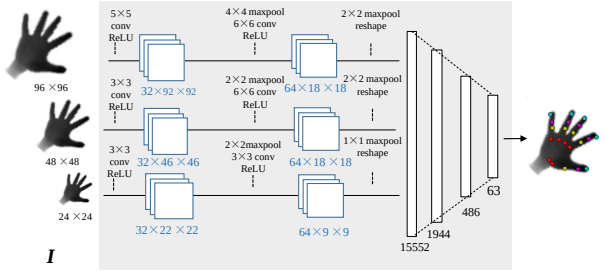

Fig. 3. Structure of the CNN model for the baseline Holi. The CNN model estimates all the 21 joints with multi-resolution input images.

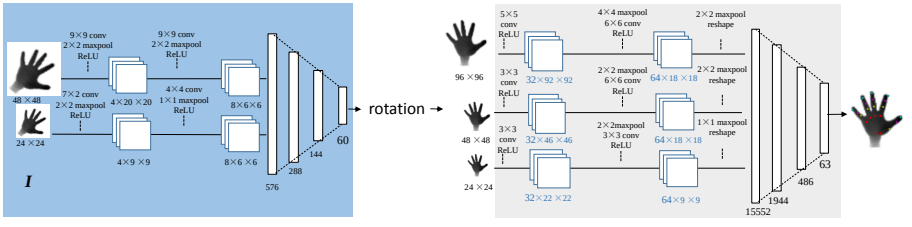

Fig. 4. Structure of the CNN models for the baseline Holi_Derot. Left: a CNN model classifies the depth image into 60 in-plane rotation bins. Right: a CNN model estimates all the 21 joints with input images rotated by the classification result.

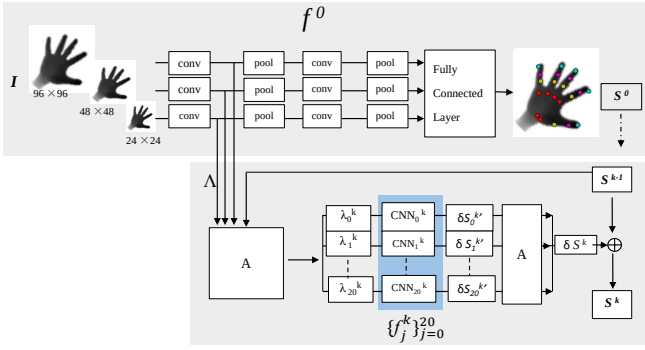

Fig. 5. Structure of the baseline Holi_SA. Top: a CNN model provides initial states for the estimation of all the 21 joints. Bottom: CNN models refine the estimation results with input and output space transformed by spatial attention module A in cascaded manner.

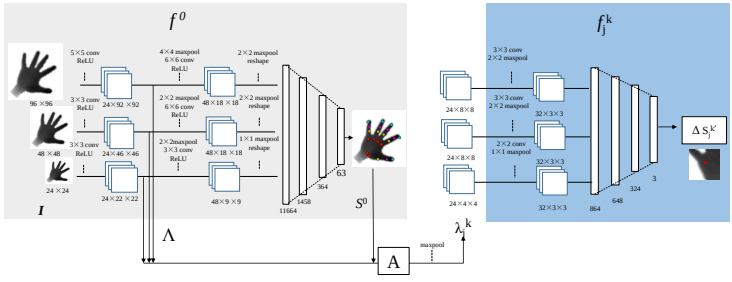

Fig. 6. Structure of the CNN models for the baseline Holi_SA in Fig. 5. Left: the CNN model in the initial stage. Right: a CNN model for a single joint j refinement in the stage k.

2 Set the Size of the Feature Maps

Here we explains the setting of the size of the feature maps λ_{lj}^k for $f_{lj}^k (k > 0)$, which is the output of the spatial attention module A after maxpooling.

According to the parameter setting for the spatial attention module A in the paper, half of the width of the output feature maps of the module A is set to be two times of the larger one of the means of the estimation errors along x and y coordinates in the initial stage. The average estimation error of a joint location along x or y coordinate is about 10 mm to 20 mm for different datasets. The size of the hands in these datasets is about 150 mm to 180 mm. Therefore, the ratio of the error range to the hand size is about 1/15 to 1/9 and accordingly, the ratio of the size of the output feature maps to the size of the input hand images is around 4/15 to 4/9. For the channel corresponding to the input images with the resolution 96×96, the size of the input feature maps for the spatial model

A is 92×92 and the size of the output feature maps is set to be about one third of the size of 92×92, i.e. 32×32. For the other two resolutions, the ratio is the same and the size of the output feature maps is 16×16 and 8×8, respectively. These output features are then maxpooled as mentioned in Sec 1.

For the input λ_{lj}^0 for the initial stage $f_{lj}^0 (l > 0)$ in Layer $l (l > 0)$, half of the width of the output feature maps of the spatial attention module A is set to be two times of the larger one of the means of the offsets along x and y coordinates for the joint lj. The offset in our experiment is around 15 mm to 30 mm. Therefore, the ratio of the offset range to the hand size is about 1/10 to 1/6 and accordingly, the ratio of the size of the output feature maps to the size of the input hand images is around 4/10 to 4/6. The ratio of the size of the output feature maps to the size of the input hand images is set 1/2 in our experiment and the size of the output feature maps for different resolution channels is set as 48×48, 24×24, 12×12.

To test the influence of the size of the feature maps to estimation result, the size of the feature maps $\lambda_{lj}^k (k > 0)$ is changed to 36×36, 18×18, 9×9 and 28×28, 14×14, 7×7 and the size of the feature maps forλ_{lj}^0 is changed to 56×56, 28×28, 14×14 and 40×40, 20×20, 10×10. The kernel size and the pool size of the CNN is adjusted a little to make the number of the units of the fully connected layers unchanged for different sizes. The fluctuation of the errors of these variations is under the 1 mm.

References

1. Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. TOG (2014)
2. Oberweger, M., Wohlhart, P., Lepetit, V.: Hands deep in deep learning for hand pose estimation. arXiv preprint arXiv:1502.06807 (2015)
3. Oberweger, M., Wohlhart, P., Lepetit, V.: Training a feedback loop for hand pose estimation. In: ICCV (2015)
4. Tang, D., Taylor, J., Kohli, P., Keskin, C., Kim, T.K., Shotton, J.: Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In: ICCV (2015)