Imperial College London

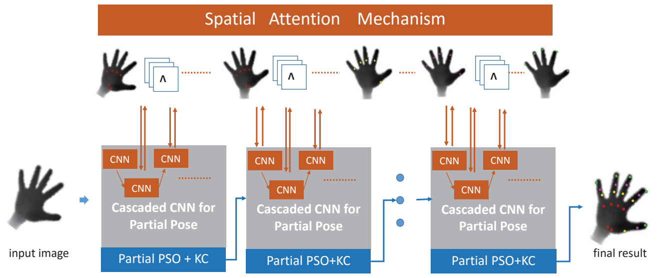
Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation

Qi Ye*, Shanxin Yuan*, Tae-Kyun Kim
{q.ye14, s.yuan14, tk.kim}@imperial.ac.uk

Motivation

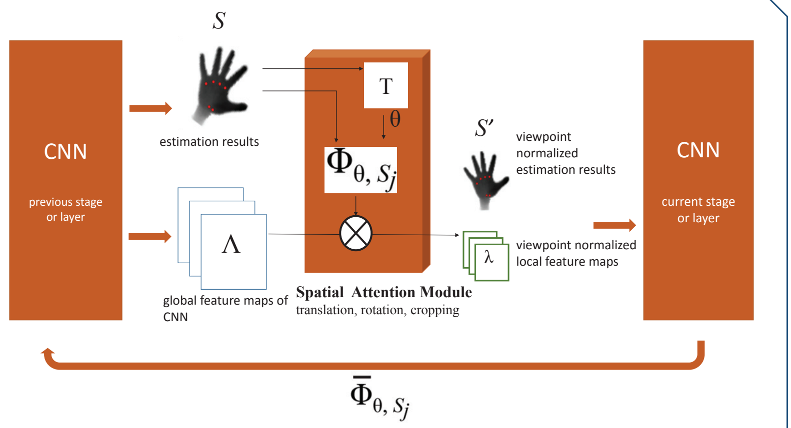
- Existing hierarchical methods mainly focus on the decomposition of the output space while the input space remains almost the same along the hierarchy.
- The spatial attention mechanism is proposed to integrate cascaded and hierarchical regression into a CNN framework by transforming both the input(and feature space) and the output space, which greatly reduces the viewpoint and articulation variations.
- Between the levels in the hierarchy, the hierarchical PSO forces the kinematic constraints to the results of the CNNs.

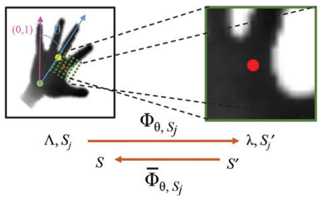
Structure



- The Spatial Attention Mechanism integrates the cascaded and hierarchical hand pose estimation into one framework.
- The hand pose is estimated layer by layer in the order of the articulation complexity, with the spatial attention module to transform the input, feature and output space.
- Within each layer, the partial pose is iteratively refined both in viewpoint and location with the spatial attention module, which leads both the feature and output space to a canonical one.
- After the refinement, the partial PSO is applied to select estimations within the hand kinematic constraints among the results of the cascaded estimation.

Spatial Attention Mechanism





$$(\lambda_j^k, S_j^{k-1'}) = \Phi_{\theta^{k-1}, S_j^{k-1}}(\Lambda, S_j^{k-1})$$

$$S_j^k = \bar{\Phi}_{\theta^{k-1}, S_j^{k-1}}(f_j^k(\lambda_j^k) + S_j^{k-1'})$$

- Spatial attention mechanism transforms the feature maps Λ and joint locations S from previous stages or layer to a new aligned space (first equation) and the transformed result is fed into the current CNN, whose estimation result is transformed back by the inverse transformation (second equation).

Kinematic Constraint

- Energy function for each layer

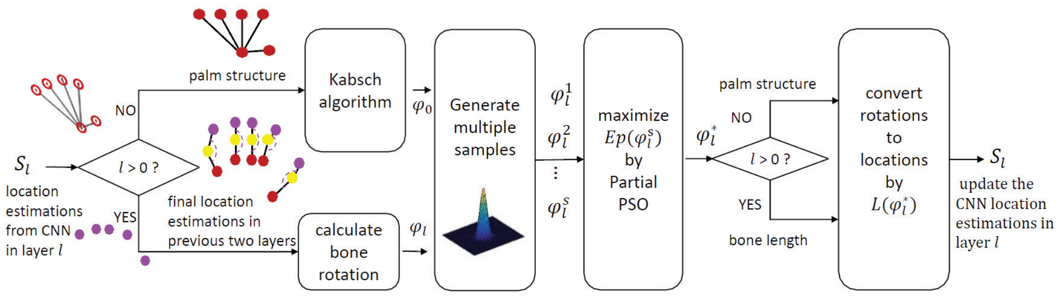
$$Ep(\varphi_l^s) = P(\varphi_l^s)Q(\varphi_l^s)$$

$$P(\varphi_l^s) \propto N(\varphi_l^s; \varphi_l, \Sigma)$$

- prior probability of the s_{th} sample *belonging to the Gaussian distribution centred on the estimation results from CNN*

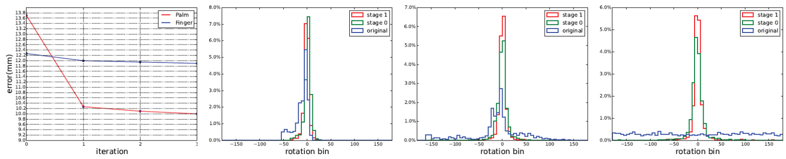
$$Q(\varphi_l^s) \propto \sum_{S_{lj}^s \in L(\varphi_l)} [B(S_{lj}^s) + D(S_{lj}^s)]$$

- force each joint to lie inside the hand silhouette and inside the depth range of a major point cloud.

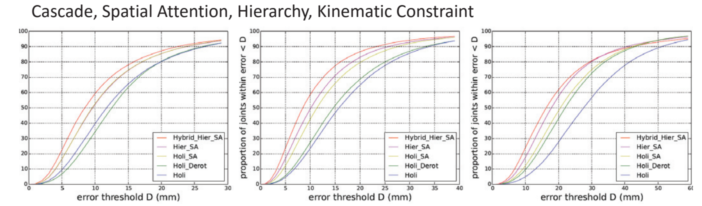


Results

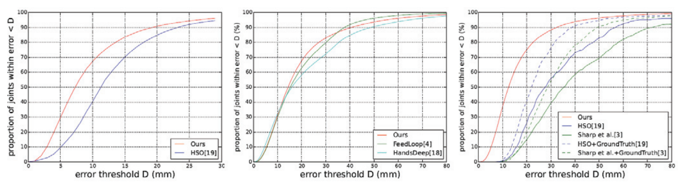
- Errors for a joint of 4 cascaded stages;
  In-plane viewpoint distribution of testing set for different stages on ICVL, NYU AND MSRC datasets



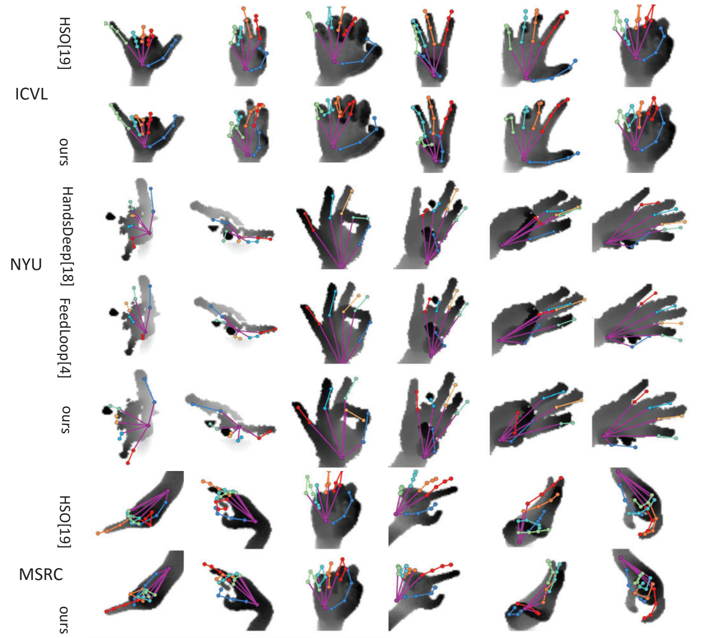
- Self comparison: demonstrate the impact of different strategies by topping them up to the baselines: Cascade, Spatial Attention, Hierarchy, Kinematic Constraint



- Comparison with state-of-the-art methods



Qualitative results



[3] Sharp, T., Keskin, C., Robertson, D., Taylor, J., Shotton, J., Leichter, D.K.C.R.I., Wei, A.V.Y., Krupka, D.F.P.K.E., Fitzgibbon, A., Izadi, S.: Accurate, robust, and flexible real-time hand tracking. In: CHI (2015)
[4] Oberweger, M., Wohlhart, P., Lepetit, V.: Training a feedback loop for hand pose estimation. In: ICCV (2015)
[18] Oberweger, M., Wohlhart, P., Lepetit, V.: Hands deep in deep learning for hand pose estimation. arXiv preprint arXiv:1502.06807 (2015)
[19] 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)

* Indicates equal contribution.