ISCAS 中国科学院软件研究所学术年会'2022
暨计算机科学国家重点实验室开放周
学术论文

基于虚拟视角选择的三维人手姿势估计

作者：程坚、万炎广、左德鑫、马翠霞、古鉴、谭平、王宏安、邓小明、张寅达

Efficient Virtual View Selection for 3D Hand Pose Estimation，**AAAI**，2022

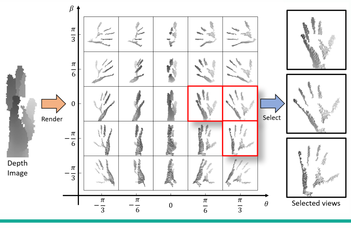
联系方式：邓小明, 13717981135, xiaoming@iscas.ac.cn

Motivation

Hand pose estimation plays a key role in many applications to support human computer interaction.

Existing methods usually use single depth as input, but estimations are not so satisfied due to viewpoint variations and self-occlusion.

The raw camera view may not be suitable for pose estimation, and finding a suitable virtual viewpoint to reproject a given single input depth may be critical for improving hand pose estimation performance.



Introduction

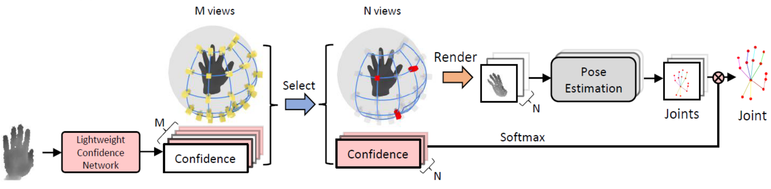
We propose a new virtual multi-view hand pose estimation method to address viewpoint variation issues on single depth.

Our main contributions:

1. We propose a novel deep learning network to predict 3D hand pose estimation from single depth;
2. We then show that the view selection can be done efficiently without sacrificing runtime via network distillation;
3. Extensive experiments demonstrate that our method achieves the state-of-the-art.

Project Webpage:

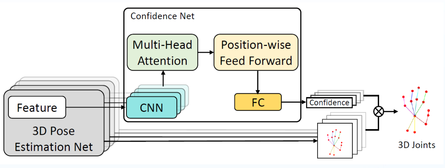
Network Architecture



Our network first generates the confidences of 25 uniformly-sampled virtual views. Then select Top-N views and render their depth. Finally, poses from these virtual views will be predicted, and we use a confidence-based fusion to get the final pose.

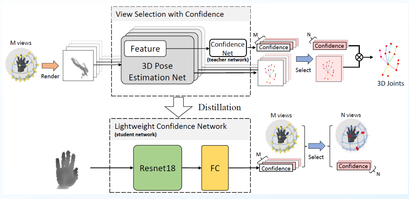
Virtual View and Confidence Generation

Our virtual view candidates are 25 uniformly-sampled views on a spherical surface.



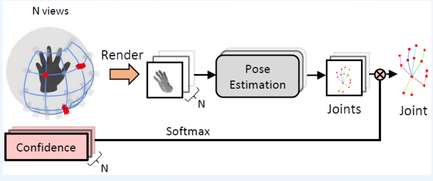
The confidence network takes the high-level feature from pose estimation network as input and evaluates the confidence of each view through multi-head attention. We use the confidence as a weight to perform a weighted average of hand pose of each view to get the final output. The confidence network is optimized by supervising the final pose.

Distillation of Confidence Network



A lightweight network (ResNet-18 followed by a fully connected layer) will be distilled, which directly takes the original depth as input and predicts the confidence of all virtual views.

Virtual Multi-view Hand Pose Estimation



Selected views will be rendered to depth, and the estimated poses from these virtual views will be predicted and fused with view confidence to get the final pose.

Comparison with SOTA

Performance on NYU, ICVL and Hands2019 Task 1

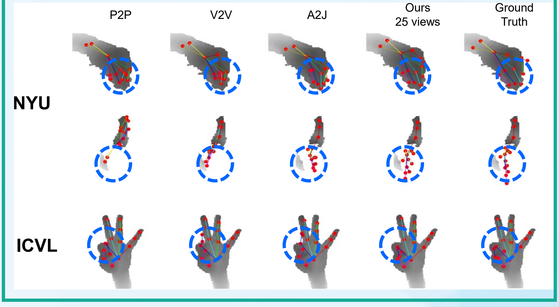
Methods	NYU	ICVL
HandPointNet (Ge et al. 2018)	10.54	6.93
DenseReg (Wan et al. 2018)	10.21	7.24
P2P (Ge, Ren, and Yuan 2018)	9.05	6.33
A2J (Xiong et al. 2019)	8.61	6.46
V2V (Moon, Chang, and Lee 2018)	8.42	6.28
AWR (Huang et al. 2020)	7.37	5.98
Ours-1view	7.34	5.16
Ours-3views	6.82	4.86
Ours-9views	6.53	4.77
Ours-15views	6.41	**4.76**
Ours-25views	**6.40**	4.79

Mean error on NYU and ICVL

Methods	Mean error (mm)
V2V (Moon, Chang, and Lee 2018)	15.57
AWR (Huang et al. 2020)	13.76
A2J (Xiong et al. 2019)	13.74
Rokid (Zhang et al. 2020)	13.66
Ours-1view	14.14
Ours-3views	13.24
Ours-9views	12.67
Ours-15views	**12.51**
Ours-25views	12.55

Mean error on Hands2019 Task1

Visualization Results



Ablation Study

Virtual multi-view, view selection and confidence-based pose fusion are all effective.

	NYU			ICVL			Hands2019-Task1		
Number of views	UNIFORM	SELECT	LIGHT	UNIFORM	SELECT	LIGHT	UNIFORM	SELECT	LIGHT
1 view	7.93	7.23	7.34	5.56	5.18	5.16	14.39	14.03	14.14
3 views	7.14	6.73	6.82	5.27	4.85	4.86	13.67	13.07	13.24
9 views	6.77	6.43	6.53	4.96	4.77	4.77	12.81	12.60	12.67
15 views	6.55	6.38	6.41	4.85	4.76	4.76	12.61	12.51	12.51
25 views	6.40	-	-	4.79	-	-	12.55	-	-

Comparison of mean joint error (in mm) using uniform sampling and view selection. “UNIFORM” denotes using uniformly sampled views. “SELECT” denotes selecting views from 25 uniformly sampled views with the “teacher” confidence network. “LIGHT” denotes selecting views from 25 uniformly sampled views with the “student” lightweight confidence network.