HandVoxNet: Deep Voxel-Based Network for 3D Hand Shape and Pose Estimation from a Single Depth Map

Jameel Malik[1,2,3] Ibrahim Abdelaziz[1,2] Ahmed Elhayek[2,4] Soshi Shimada[5]
Sk Aziz Ali[1,2] Vladislav Golyanik[5] Christian Theobalt[5] Didier Stricker[1,2]

[1] TECHNISCHE UNIVERSITÄT KAISERSLAUTERN [2] DFKI [3] [4] [5] max planck institut informatik

Motivation

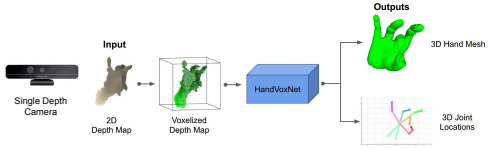
Accurate 3D hand shape and pose estimation has many applications such as **animation**, **signing in the air** and **handling virtual objects** in VR/AR

Major Challenges

- Varying hand shapes
- High DOF, occlusion and self-similarity
- Annotating real images for shape is hard

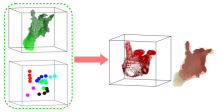
Contributions

The first method, called **HandVoxNet**, which simultaneously estimates 3D hand shape and pose using **3D convolutions**

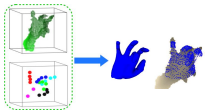


Novelties

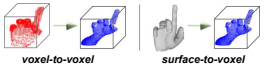
voxel-to-voxel 3D CNN network (voxelized shape)



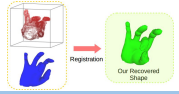
voxel-to-surface 3D CNN network (shape surface)



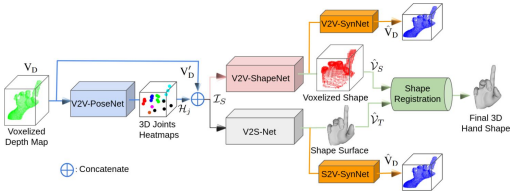
Voxelized **depth map synthesizers**



Hand shape registration (DispVoxNet and NRGA)



Method



Ablation Study

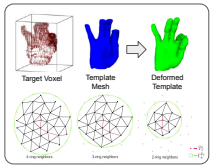
Methods	3D $\mathcal{V}$ Err. (mm)
V2S-Net (w/o $\mathcal{H}_J$)	8.78
V2S-Net (w/o $\mathcal{V}_D'$)	3.54
V2S-Net (with $\mathcal{H}_J \oplus \mathcal{V}_D'$)	**3.36**
Methods	**3D $\mathcal{S}$ Err.**
V2V-ShapeNet (w/o $\mathcal{H}_J$)	0.007
V2V-ShapeNet (w/o $\mathcal{V}_D'$)	0.016
V2V-ShapeNet (with $\mathcal{H}_J \oplus \mathcal{V}_D'$)	**0.005**

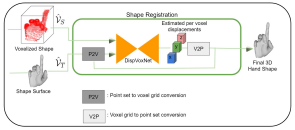
Loss Function

$$\mathcal{L}_T = \mathcal{L}_{\mathcal{H}} \quad \text{3D heatmaps of joints}$$
$$+\mathbb{1}\mathcal{L}_{\mathcal{V}_S} \quad \text{Voxelized shape (BCE)}$$
$$+\mathbb{1}\mathcal{L}_{\mathcal{V}_T} \quad \text{Shape surface (Euclidean)}$$
$$+\ \mathcal{L}^{\mathcal{V}}_{\mathcal{V}_D} \quad \text{Voxelized depth } (\hat{\mathcal{V}}_S \rightarrow \mathcal{V}_D')$$
$$+\ \mathcal{L}^{\mathcal{S}}_{\mathcal{V}_D} \quad \text{Voxelized depth } (\hat{\mathcal{V}}_T \rightarrow \mathcal{V}_D')$$

Shape Registration Methods

NRGA, 3DV'18 ⟷ **DispVoxNet, 3DV'19**





Network Training

- **Synthetic Data** (Fully Labelled)
  - V2V-PoseNet, V2V-ShapeNet and V2S-Net are separately trained with full supervision of pose and shape.
  - The networks are put together, and then fined-tuned in an end-to-end manner.
- **Combined Real and Synthetic Data**
  - V2V-SynNet and S2V-SynNet act as a source of weak-supervision.
  - $\mathbb{1}$ is 1 for synthetic and 0 for real data. Backprops of V2V-ShapeNet and V2S-Net are disabled for real data.

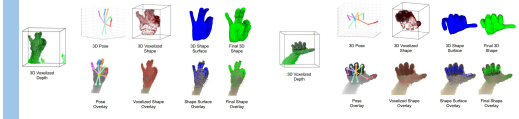
3D Data Augmentation

$$[\hat{i}, \hat{j}, \hat{k}]^{\mathrm{T}} = [\mathrm{Rot}_x(\theta_x)] \quad [-40°, +40°]$$
$$\times\ [\mathrm{Rot}_y(\theta_y)] \quad [-40°, +40°]$$
$$\times\ [\mathrm{Rot}_z(\theta_z)][i, j, k]^{\mathrm{T}} \quad [-120°, +120°]$$

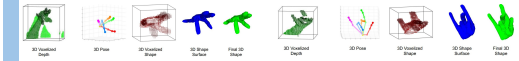
- Scaling $[+0.8, +1.2]$
- Translation $[-8, +8]$

Results

NYU dataset, TOG'14



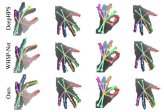
BigHand2.2M dataset, CVPR'17



Comparisons to DeepHPS, 3DV'18, and WHSP-Net, Sensors'19 on NYU

3D Shape

3D Pose



Quantitative Comparison to V2V-PoseNet, CVPR'18

Dataset	Method	3D $\mathcal{J}$ Err. (mm)
NYU	V2V-PoseNet	9.22
	V2V-PoseNet (our 3D augm.)	**8.72**
BigHand2.2M	V2V-PoseNet	9.95
	V2V-PoseNet (our 3D augm.)	**9.27**

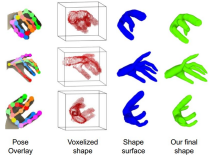
SynHand5M dataset, 3DV'18

Methods	3D $\mathcal{J}$ Err. (mm)
DeepHPS	6.30
WHSP-Net	4.32
V2V-PoseNet	3.81
our HandVoxNet (full method)	**3.75**

3D pose estimation results.

Methods	3D $\mathcal{V}$ Err. (mm)
DeepHPS	11.8
WHSP-Net	5.12
ours (w/o synthesizers)	2.92
ours (with synthesizers)	**2.67**

3D shape estimation results.



This work was funded by:

- German Federal Ministry of Education and Research as part of the project VIDETE (grant number 01IW18002).
- ERC Consolidator Grant 770784.