HAND SENSING FOR AUGMENTED INTERACTION

Junsong Yuan

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences
Outline

• Hand pose estimation and augmented interaction via depth cameras
• Hand pose estimation and augmented interaction via RGB cameras
Recent progress

The recent several years have witnessed a surging market of depth cameras and wearable devices.
Hand sensing from a depth camera
Hand MoCap: Problem Description

**Input:** a depth image containing a human hand

**Output:** estimated 3D hand joint locations (in total 21 joints) which represent the hand pose
Virtual Reality
Augmented Reality

Hui Liang, Junsong Yuan, Daniel Thalmann, IEEE Trans. on Cybernetics 2019
Challenges

- Hand pose has high degree of freedom
Challenges

• View-point variations

Same gesture from different view
Challenges

• Self-occlusions
Multi-view CNNs based Method

Multi-view Projection

• The pixel values on projection images represent the normalized projection distances of 3D points.
Architecture of CNNs

- The network generates $K$ heat-maps with the size of $18 \times 18$ pixels. All of the six views have the same network architecture and the same architectural hyperparameters.
Multi-view Fusion

Objective: estimate $K$ objective hand joint 3D locations

$$\Phi = \{\phi_k\}_{k=1}^{K} \in \Lambda$$

Maximum a posteriori estimation

$$\Phi^* = \arg \max_{\Phi} P(\Phi | I_1, I_2, \cdots, I_N)$$  \hspace{1cm} \text{posterior probability}$$

$$= \arg \max_{\Phi} P(I_1, I_2, \cdots, I_N | \Phi)$$  \hspace{1cm} \text{maximum likelihood}$$

$$= \arg \max_{\Phi} \prod_{n=1}^{N} P(I_n | \Phi)$$  \hspace{1cm} (assume equal a priori probability)$$

$$= \arg \max_{\Phi} \prod_{n=1}^{N} P(\Phi | I_n)$$  \hspace{1cm} (assume conditional independence)$$

$$= \arg \max_{\Phi} \prod_{k=1}^{K} \prod_{n=1}^{N} P(\phi_k | I_n)$$  \hspace{1cm} \text{related with heat-map}$$

s.t. $\Phi \in \Omega$, constrained to a low dimensional subspace in order to resolve ambiguous joint estimations
Multi-view Fusion

\[ Q(\phi_k = p) = \prod_{n=1}^{N} P(\phi_{kn} = p_n | I_n) \]
From 2D convolution to 3D convolution?

2D convolution for depth image
From 2D convolution to 3D convolution?

2D convolution for depth image

2D-CNN

Heat-maps

3D Joint Locations

Depth Image

But 3D points are sparse data

Dense 3D convolution on sparse point clouds will fail

Depth Image

3D Points

depth image can be transferred to 3D points
From 2D convolution to 3D convolution?

2D convolution for depth image

Transfer sparse 3D points to dense volumetric representation
• L. Ge, H. Liang, J. Yuan, and D. Thalmann. 3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images. CVPR 2017 and TPAMI 2019.
Volumetric Representation

- Projective Directional Truncated Signed Distance Function (D-TSDF) for volumetric representation

Network Architecture

- Input: three volumes of the projective D-TSDF
- Output: a column vector containing $3 \times K$ elements corresponding to the $K$ 3D hand joint relative locations in the volume.
- Three 3D convolutional layers and three fully-connected layers
**Network Architecture**

- **Input**: three volumes of the projective D-TSDF
- **Output**: a column vector containing $3 \times K$ elements corresponding to the $K$ 3D hand joint relative locations in the volume.
Patterns learned in 3D shallow network

- Neurons in layer 1 (L1) can capture local structures, such as corners and edges;
- neurons in layer 2 (L2) can capture structures of hand part, such as fingers;
- neurons in layer 3 (L3) can capture global structures of hand.
3D Data Augmentation

Introducing variations of training data
Data Augmentation Can Help Training
Test on MSRA hand pose dataset

Test on MSRA hand pose dataset

Test on NYU hand pose dataset

3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images

Liuhaok Ge, Hui Liang, Junsong Yuan, Daniel Thalmann
Institute for Media Innovation
Nanyang Technological University
Can we process 3D point cloud directly instead of 3D convolution?

3D convolution for depth image

Process 3D point cloud directly?
Hand PointNet

A **point cloud** based hand pose estimation approach by holistically regressing the 3D hand pose.
Further improvement

Estimate the point-wise closeness and offset directions to hand joints from the input point cloud using a stacked point-to-point regression PointNet, which is able to capture local evidence for estimating accurate 3D hand pose.
Hand PointNet: 3D Hand Pose Estimation using Point Sets

Liu Hao Ge\textsuperscript{1}, Yujun Cai\textsuperscript{1}, Junwu Weng\textsuperscript{1}, Junsong Yuan\textsuperscript{2}

\textsuperscript{1}Nanyang Technological University
\textsuperscript{2}State University of New York at Buffalo
3D hand pose estimation: Can we use RGB camera instead of depth camera?
Monocular RGB-based Approach

From 2D images to 3D skeleton results
Challenges: difficult to obtain 3D labeled data

For Real Dataset:

- annotate accurate 3D hand pose is difficult
- Multi-view annotation method is labor-costing
- Reconstructed 3D labels may not be perfect
Using synthetic data for machine learning?

synthetic dataset for hand pose [Zimmermann et al. ICCV 2017]

Synthetic data can provide accurate 3D annotations while quite different from real ones
Weakly Supervised Learning

Y Cai, L Ge, J Cai, J Yuan, Weakly-supervised 3d hand pose estimation from monocular RGB images, ECCV’18
From 3D hand pose estimation to joint 3D hand pose and shape estimation
Joint hand pose and shape estimation

Input Image

2D/3D Locations of Hand Joints
Challenges

• High dimensionality of the output space (3D mesh)
  ➢ We propose a novel Graph CNN-based approach to generate 3D hand mesh vertices in a graph

• Lack of ground truth 3D hand mesh training data for real-world images
  ➢ We propose a novel weakly-supervised method by leveraging depth map as a weak supervision for 3D mesh generation
Method – Overview

If we cannot solve a simple problem, try a complex one
Method – Graph CNN

Chebyshev Spectral Graph CNN [1]

Method – Graph CNN

Graph CNN for Mesh Generation

Latent Feature (4096) → FC (512), ReLU → FC (5120) → Reshape → 80×64 → Upsampling (×4) → Graph Conv (64), BN, ReLU → Graph Conv (32), BN, ReLU → Upsampling (×4) → Graph Conv (32), BN, ReLU → Graph Conv (3) → 1280×3
Method – Graph CNN

Graph CNN for Mesh Generation

(a) Graph Coarsening

(b) Example of Feature Upsampling
Method − Training

Fully-supervised Training on Synthetic Dataset

Loss Function \( \mathcal{L}_{fully} = \lambda_H \mathcal{L}_H + \lambda_M \mathcal{L}_M + \lambda_J \mathcal{L}_J \)

Mesh Loss \( \mathcal{L}_M = \lambda_v \mathcal{L}_v + \lambda_n \mathcal{L}_n + \lambda_e \mathcal{L}_e + \lambda_l \mathcal{L}_l \)

- Heat-map Loss
- Mesh Loss
- 3D Pose Loss
- Vertex Loss
- Normal Loss
- Edge Loss
- Laplacian Loss
Method – Training

Weakly-supervised Finetuning on Real-world Dataset

Loss Function

\[ \mathcal{L}_{weakly} = \lambda_{H} \mathcal{L}_{H} + \lambda_{D} \mathcal{L}_{D} + \lambda_{PN} \mathcal{L}_{PN} \]

- Heat-map Loss
- Depth Map Loss
- Pseudo-GT Mesh Loss
A Large Synthetic Dataset for Training and Validation
(375,000 RGB images with hand mesh and pose annotations)

Synthetic Dataset Creation
3D Hand Shape and Pose Dataset

• A Large Synthetic Dataset for Training and Validation (375,000 hand RGB images)

• A Real-world Dataset for Testing (583 hand RGB images)
Experiments
Evaluation of 3D Hand Mesh Reconstruction

<table>
<thead>
<tr>
<th>Error (mm)</th>
<th>Normal</th>
<th>Edge</th>
<th>Laplacian</th>
<th>3D Pose</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh error</td>
<td>8.34</td>
<td>9.09</td>
<td>8.63</td>
<td>9.04</td>
<td>7.95</td>
</tr>
<tr>
<td>Pose error</td>
<td>8.30</td>
<td>9.06</td>
<td>8.55</td>
<td>9.24</td>
<td>8.03</td>
</tr>
</tbody>
</table>

Ablation study by eliminating different loss terms from our fully-supervised training loss.
Experiments

Evaluation of 3D Hand Mesh Reconstruction

<table>
<thead>
<tr>
<th>Mesh error (mm)</th>
<th>MANO-based</th>
<th>Direct LBS</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our synthetic dataset</td>
<td>12.12</td>
<td>10.32</td>
<td>8.01</td>
</tr>
<tr>
<td>Our real-world dataset</td>
<td>20.86</td>
<td>13.33</td>
<td>12.72</td>
</tr>
</tbody>
</table>

Comparison with direct Linear Blend Skinning (LBS) method and MANO-based method.
Experiments

Evaluation of 3D Hand Pose Estimation

Comparisons with state-of-the-art methods on STB dataset
3D mesh + 3D pose estimation
3D Hand Shape and Pose Estimation from a Single RGB Image

Input

2D/3D locations of hand joints

3D hand mesh

PaperID 387
Summary

• Hand Sensing for Augmented interactions
  • Hands are important tools for interactions and communications
  • Hand sensing from depth camera and optical camera
  • If we cannot solve a simple problem, try a complex one

• Graphics is more than rendering
  • Graphics synthesised data play important role for AI
  • We want creations that look both real and smart
Thank you!