Robust 3D Hand Pose Estimation in Single Depth Images: from Single-View CNN to Multi-View CNNs

Lihao Ge1, Hui Liang1,2, Junsong Yuan2, and Daniel Thalmann1

1Institute for Media Innovation, Nanyang Technological University, Singapore
2School of EEE, Nanyang Technological University, Singapore

Accepted to CVPR’16

Presented by: Lihao Ge
April 13, 2016
Motivation

The recent several years have witnessed a surging market of depth cameras and wearable devices.
Motivation

3D hand pose estimation plays a significant role in human-computer interaction, such as virtual/augmented reality applications.
Related Work

Model-driven approaches [1, 2]:
- Find the optimal hand pose parameters via fitting a deformable 3D hand model to input image observations
- Need to pre-define hand size and motion constraints; sensitive to initialization

Data-driven approaches [3, 4, 5, 6]:
- Train discriminative models to map image features to hand pose parameters
  - Random Forest (Tang et al. 2014 [3], Sun et al. 2015 [4])
  - CNN (Tompson et al. 2014 [5], Oberweger et al. 2015 [6])
- Do not require model calibration and robust to poor initialization
Related Work

- We focus on **CNN-based data-driven methods** in this work.
- Tompson et al. [5] proposed to map input image to a set of **heat-maps** which represent the probability distributions of joint positions in the image and recover the 3D joint locations from the depth image with model fitting.
Related Work

• In this method, the depth of a hand joint is taken as the corresponding depth value at the estimated 2D position.

• This may result in large depth estimation errors even if the estimated 2D position is only slightly deviated from the true joint position.
Related Work

- In case of **ambiguous estimations**, the single view CNN cannot well differentiate among multiple hotspots in the heat-map, in which only one could correspond to the true joint.

- In essence, in this method, the heat-map only provides 2D information of the hand joint and the **depth information is not fully utilized**.
**System Overview**

- In this work, we propose a novel 3D regression method using **multi-view CNNs** that can better exploit depth cues to recover fully 3D information of hand joints without model fitting.
System Overview

Depth Image → 3D Points → Projections → Convolutional Networks → Heat-maps → Multi-view Fusion → 3D Joints
The pixel values on projection images represent the normalized projection distances of 3D points.
• The network generates 21 heat-maps with the size of 18x18 pixels. All of the three views have the same network architecture and the same architectural parameters.
Multi-view Fusion

• We estimate the hand joint locations $\Phi$ by applying the \textbf{MAP (maximum a posterior) estimator} on the basis of projections $I_{xy}, I_{yz}$ and $I_{zx}$.

• Given $(I_D, \Phi)$, we assume that the three projections $I_{xy}, I_{yz}$ and $I_{zx}$ are independent, conditioned on the joint locations $\Phi$.

$$
\Phi^* = \arg \max_{\Phi} P(\Phi | I_{xy}, I_{yz}, I_{zx})
= \arg \max_{\Phi} P(I_{xy}, I_{yz}, I_{zx} | \Phi)
= \arg \max_{\Phi} P(I_{xy} | \Phi) P(I_{yz} | \Phi) P(I_{zx} | \Phi)
= \arg \max_{\Phi} P(\Phi | I_{xy}) P(\Phi | I_{yz}) P(\Phi | I_{zx})
$$

$s.t. \Phi \in \Omega$

$\Phi$ is constrained to a \textbf{low dimensional subspace} $\Omega$ in order to resolve ambiguous joint estimations.
Multi-view Fusion

\[
 \Phi^* = \arg \max_\Phi \sum_k \log Q(\phi_{kx}, \phi_{ky}, \phi_{kz})
\]

\[
 = \arg \max_\Phi \sum_k \log \mathcal{N}(\mu_k, \Sigma_k)
\]

\[
 = \arg \min_\Phi \sum_k (\phi_k - \mu_k)^T \Sigma_k^{-1} (\phi_k - \mu_k)
\]

s.t. \( \Phi = \sum_{m=1}^M \alpha_m e_m + u \)

\( Q(\phi_{kx}, \phi_{ky}, \phi_{kz}) \) denotes the product of probabilities \( P(\phi_{kx}, \phi_{ky} \mid I_{xy}), P(\phi_{ky}, \phi_{kz} \mid I_{yz}), P(\phi_{kz}, \phi_{kx} \mid I_{zx}), \) which can be obtained from three views’ heat-maps of joint \( k \).
Experimental Results

Test on MSRA hand pose dataset [4]:

![Graph showing experimental results](image-url)
Experimental Results

1st line: single view regression,
2nd line: multi-view regression with coarse fusion,
3rd line: multi-view regression with fine fusion
References


Thank You!