Training a Feedback Loop for Hand Pose Estimation

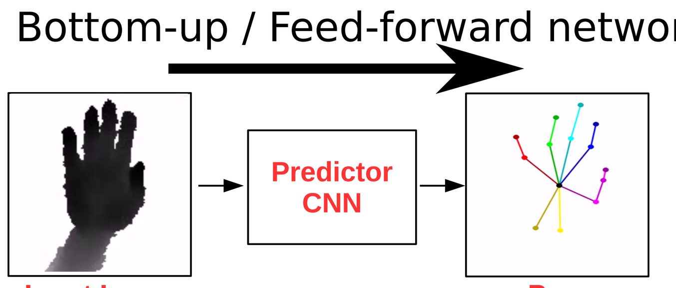
Markus Oberweger, Paul Wohlhart and Vincent Lepetit

Graz University of Technology, Institute for Computer Vision and Graphics, cvarlab.icg.tugraz.at

Motivation

- Goal: Accurate 3D hand pose estimation from single depth image
- Straightforward approach: Predictor for 3D pose

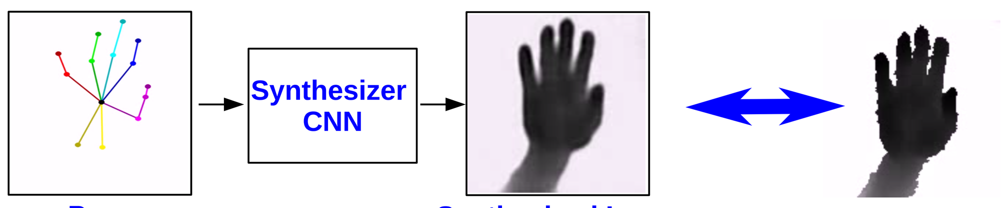
Bottom-up / Feed-forward network



Many neurons in the visual cortex provide *feedback*

- Adding top-down information with a synthesizer

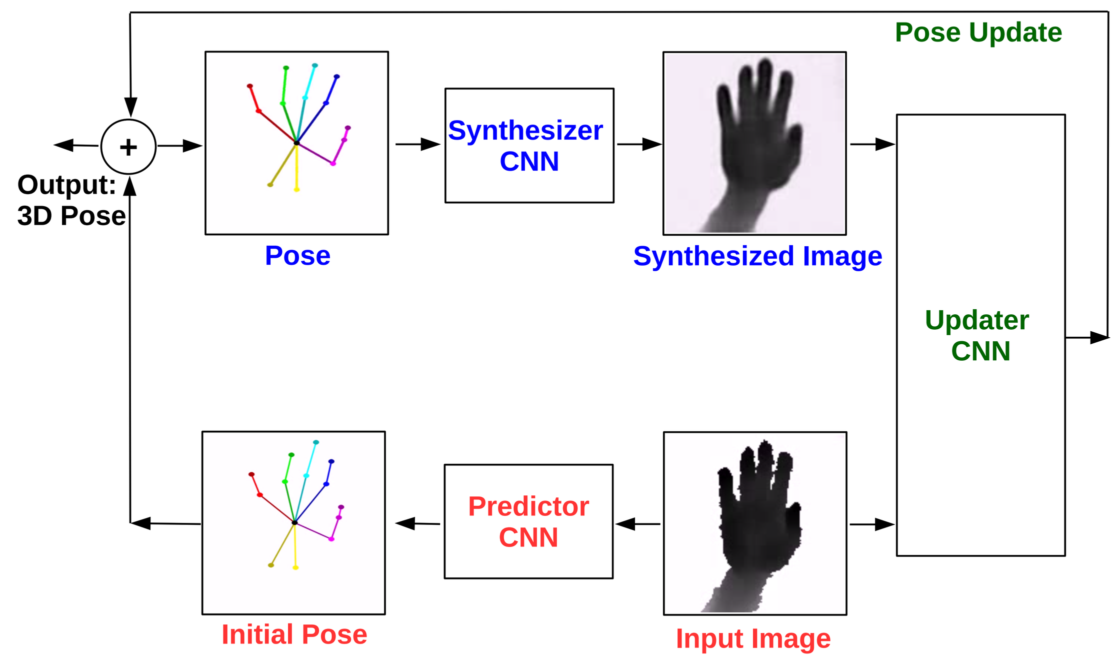
Synthesizes depth image from 3D pose



No 3D hand model required!

Difficult to estimate pose by minimizing difference between synthesized and input image!

- Our approach: Train an updater to predict updates for the initial 3D pose of a predictor using a learned model of the hand



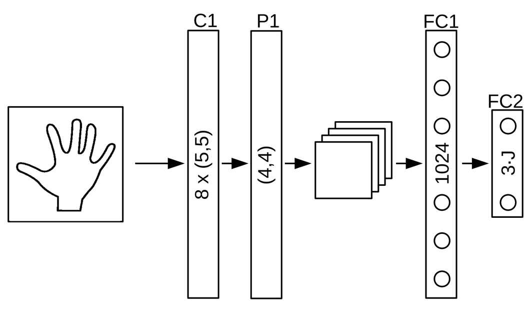
- All components are implemented as Convolutional Neural Networks and trained from data.

Method

- Training a predictor CNN to predict initial 3D pose $\mathbf{p}$ from depth image $\mathcal{D}$

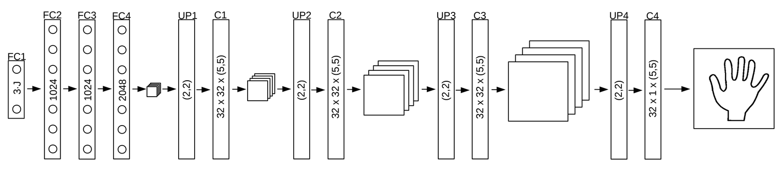
$$\arg\min_\Phi \sum_{(\mathcal{D},\mathbf{p})\in\mathcal{T}} \|\text{pred}_\Phi(\mathcal{D})-\mathbf{p}\|_2^2+\gamma\|\Phi\|_2^2$$

CNN architecture

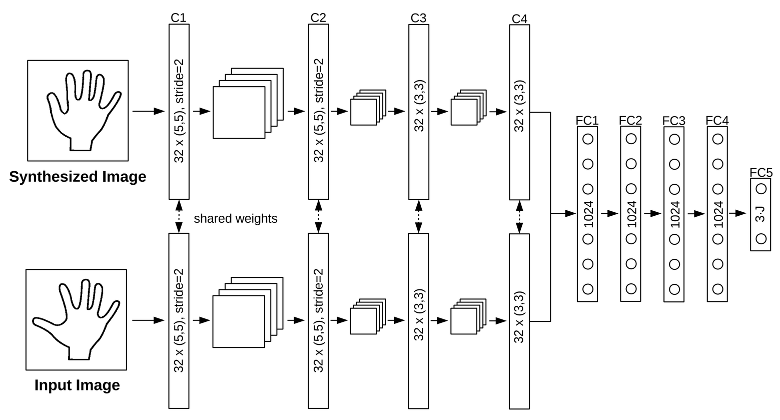


- Training a synthesizer CNN to generate depth image $\text{synth}(\mathbf{p})$ from 3D pose $\mathbf{p}$

$$\arg\min_\Theta \sum_{(\mathcal{D},\mathbf{p})\in\mathcal{T}} \frac{1}{|\mathcal{D}|}\|\text{synth}_\Theta(\mathbf{p})-\mathcal{D}\|_2^2$$



- Training an updater CNN to predict 3D pose update from two depth images



Predict update in single step: Difficult!

Iterative updates: Tractable!

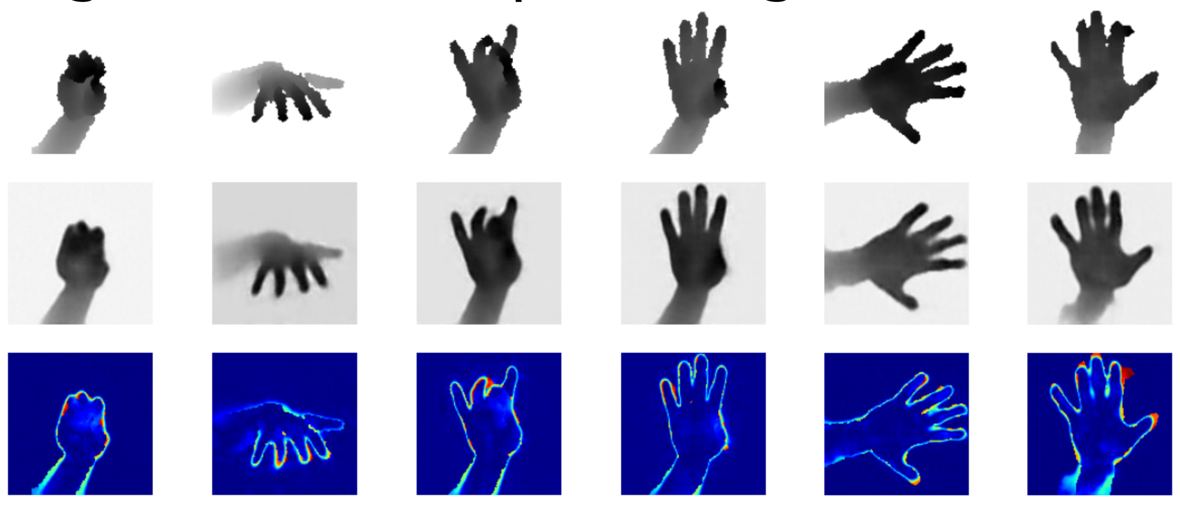
$$\arg\min_\Omega \sum_{(\mathcal{D},\mathbf{p})\in\mathcal{T}}\sum_{\mathbf{p}'\in\mathcal{T}_\mathcal{D}} \max(0, \|\mathbf{p}'+\text{updater}_\Omega(\mathcal{D},\text{synth}(\mathbf{p}'))-\mathbf{p}\|_2-\lambda\|\mathbf{p}'-\mathbf{p}\|_2)$$

- Effectively augment the training data $\mathcal{T}_\mathcal{D}$
  - Add Gaussian noise to 3D poses
  - Predictions on training data
  - Samples from error distribution of predicted updates

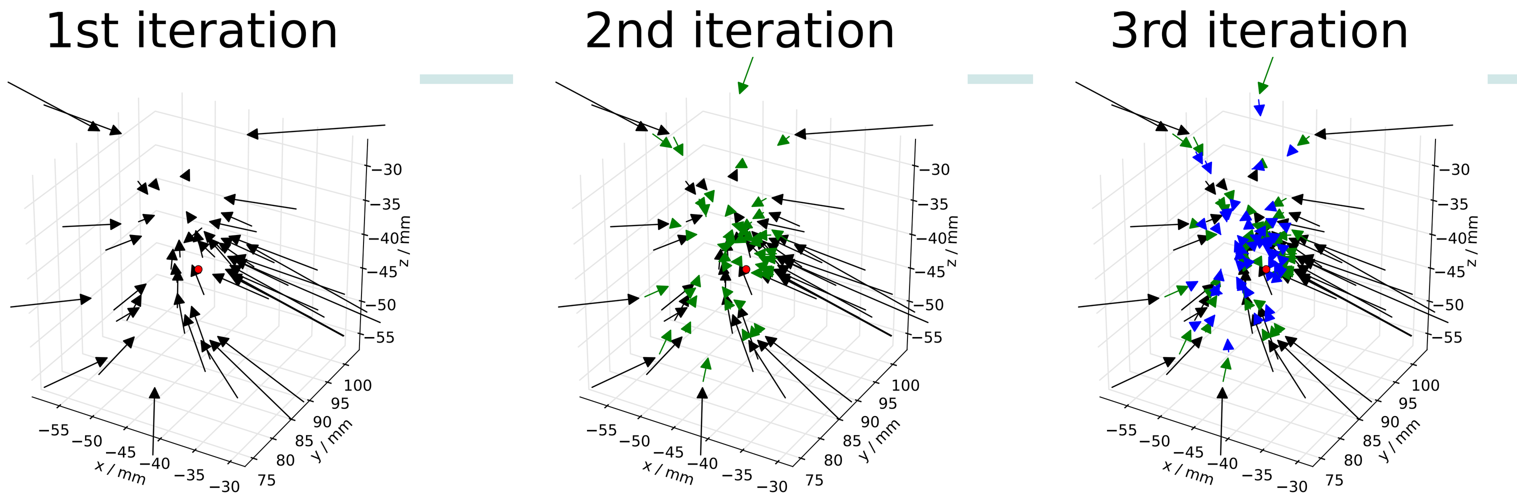
More details about the training can be found in the paper.

Results

- Synthesizing accurate depth images from 3D pose



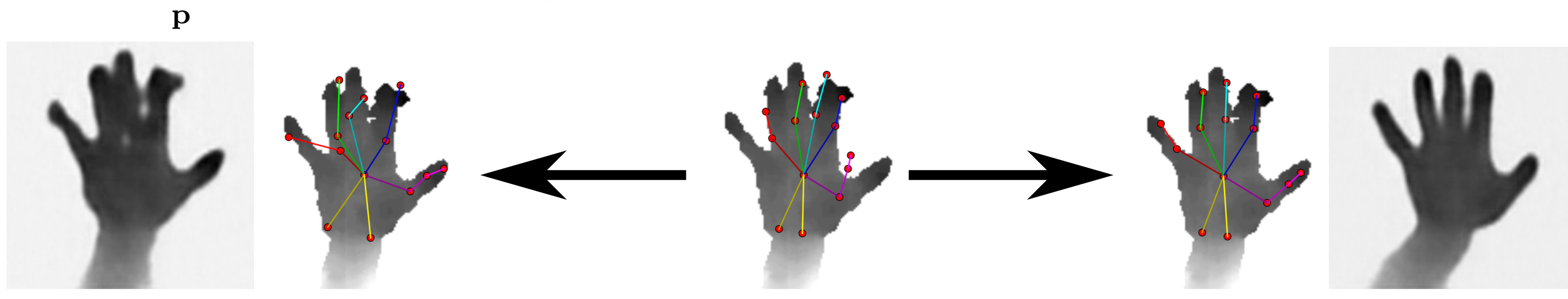
- Predicted pose updates



- Update mechanism is robust to image noise

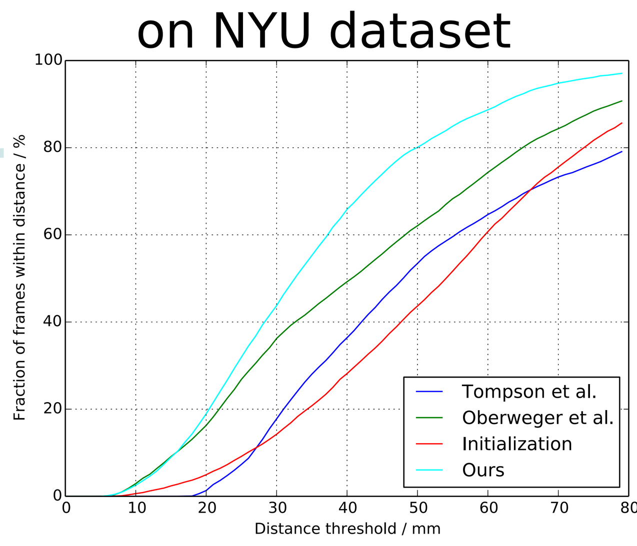
Minimize image difference:
$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}}\|\mathcal{D}-\text{synth}(\mathbf{p})\|_2^2$

Our approach:



- Quantitative evaluation of predicted 3D pose

Comparison with SOTA on NYU dataset



Different initialization