Training a Feedback Loop for Hand Pose Estimation

Markus Oberweger, Paul Wohlhart and Vincent Lepetit

Institute for Computer Vision and Graphics
Graz University of Technology
Motivation

• Important for Human Computer Interaction, Augmented Reality

• Hot topic: [Tang et al., ICCV’15], [Tzionas&Gall, ICCV’15], [Li et al., ICCV’15], [Supančič et al., ICCV’15], [Choi et al., ICCV’15], [Rogez et al., ICCV’15], [Xu&Cheng, ICCV’13], [Tompson et al., ToG’14], [Tang et al., ICCV’13], [Sridhar et al., ICCV’13], [Sharp et al., CHI’15], [Qian et al., CVPR’14], [Oikonomidis et al., ICCV’11], [Melax et al., GIC’13], [Kuznetsova et al., ICCV’13], [Keskin et al., ICCV’11], [de La Gorce et al., PAMI’11], [Ballan et al., ECCV’12], Leap Motion

• Our goal: Accurate 3D pose from single depth image
Challenges

- Self-occlusions
- Self-similarity
- Many degrees-of-freedom
- Noisy input
A Predictor

Many neurons in the visual cortex provide *feedback*

---

Initial Pose → Predictor CNN → Input Image

Feed-forward network
Starting a Feedback Loop:
Training a CNN to synthesize depth images from pose

Minimizing the difference between the two images to estimate the pose?

Does not work well in practice!

No need for a 3D model

[de la Gorce et al.]
Iterating the Feedback Loop

Training a CNN to predict updates for the pose
The Predictor

• Simple and very fast CNN for pose initialization

\[
\arg \min_\Phi \sum_{\mathcal{T}} \| \text{predictor}_\Phi(\text{depth}) - \text{pose} \|_2^2 + \gamma \| \Phi \|_2^2
\]
The Synthesizer

[Zeiler et al., ICCV'11; Dosovitskiy et al., CVPR'15]
The Synthesizer

\[
\arg\min_\Theta \sum_{\mathcal{T}} \|\text{synthesizer}_\Theta(\text{pose}) - \text{depth}\|_2^2
\]
Directly using the Synthesizer for Optimization?

\[
\text{pose} = \underset{\text{pose}}{\arg \min} \left\| \text{depth} - \text{synthesizer}(\text{pose}) \right\|_2^2
\]
The Updater

Synthesized Image

Input Image

32 x (5,5), stride=2

shared weights

32 x (5,5), stride=2

32 x (5,5), stride=2

32 x (3,3)

32 x (3,3)

32 x (3,3)

32 x (3,3)

FC1

FC2

FC3

FC4

FC5

3D Pose update
The Updater

- Predict updates to increase accuracy of pose

Difficult!

Feasible!

- Get closer

$$\| p + \text{updater}(\text{depth, synthesizer}(p)) - p_{GT} \|_2 < \lambda \| p - p_{GT} \|_2$$
The Updater

• Training is actually more complex...

\[
\arg \min_{\Omega} \sum_{(D,p) \in T} \sum_{p' \in T_D} \max(0, \|p' + \text{updater}_\Omega(D, \text{synth}(p')) - p\|_2 - \lambda \|p' - p\|_2)
\]
Results for Simple Initialization

Fraction of frames within distance / %

Predictor 1

Predictor 1 + Ours

Distance threshold / mm
Better Initialization: Better Results

Fraction of frames within distance / %

Distance threshold / mm

+20%@30mm
New SOTA on NYU Dataset

[Tompson et al., ToG’14]
Our Results on NYU Dataset
Conclusion

• Everything is learned: No hand-crafted 3D model and similarity measure needed!

• Trained feedback loop provides updates for pose

• Predicting updates much more robust than minimizing image difference

• Efficient implementation possible (~400fps on GPU)

• General approach, applicable to other pose estimation problems
Thanks! Questions?

{oberweger, wohlhart, lepetit}@icg.tugraz.at