**Efficiently Creating 3D Training Data for Fine Hand Pose Estimation**

Markus Oberweger, Gernot Riegler, Paul Wohlhart and Vincent Lepetit

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

**Motivation**

- SOTA hand pose estimation methods are data-driven, but how do we get accurate data?
- ICVL [3] dataset
- MSRA [2] dataset

**Method**

- **Automatic reference frame selection**
  - Select subset of frames that require user annotation
  - Compared to regular sampling: 50% less user intervention, 15% higher accuracy
  - Submodular optimization:
    - Select minimal set of reference frames that optimally cover pose space
    - Each frame increases cover
    - Exact solution is NP-hard
    - Greedy and fast algorithm often provides exact solution [5]

- **3D locations for reference frames**
  - User provides: 2D locations, joint visibility, and depth order constraints
  - Optimize for 3D locations such that:
    - Reprojection of 3D locations close to 2D user annotations
    - Visible joints in range of observed depth values
    - Hidden joints not in front of observed depth values
    - Depth order constraints of parent joints fulfilled
    - Skeleton constrained by bone length

- **Automatic inference of remaining frames**
  - Select closest pair of initialized frame and not initialized frame
  - Initialize 3D locations with closest and align with SIFTFlow [4]
  - Optimize for 3D locations:
    - Maximize similarity of joint appearance in depth map between initialized and not initialized frame
    - Skeleton constrained by bone length

- **Global optimization for all 3D locations**
  - Maximize similarity of joint appearance in depth map between reference and non-reference frame
  - Enforce temporal smoothness
  - Ensure consistency with 2D user annotations
  - Skeleton constrained by bone length

**Code and dataset are available online**

**Results**

- **Efficient reference frame selection**
  - Frames closer to reference in 3D → easier propagation to other frames
  - Close to exact solution
  - Fast runtime
  - Scales well to large datasets

- **Accurate 3D locations for reference frames**
  - Intuitive and fast to annotate (45s per frame)
  - Robust to annotation noise
  - Only minor decrease in accuracy for a fraction of annotations

- **Accuracy of full pipeline for 10% reference frames**
  - 3D locations of closest reference frames
  - Gain of SIFTFlow small, but useful for joint appearance term

**Method Overview**

- Training SOTA pose estimator [1] with better annotations
  - Using our annotations

**Goal: Accurate 3D training data for single view**

- Reduce time spent on annotations by a factor of 10
- We provide a new dataset for egocentric 3D hand pose estimation

**Code and dataset are available online**

**Using our annotations**

Using annotations of [2]

**Using annotations of [2]**

**Using our annotations**

- ICVL [3] dataset
- MSRA [2] dataset

**Our annotation**

**Results**

- Gain of SIFTFlow small, but useful for joint appearance term
- 3D locations of closest reference frames
- Gain of SIFTFlow small, but useful for joint appearance term

**More details can be found in the paper.**