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

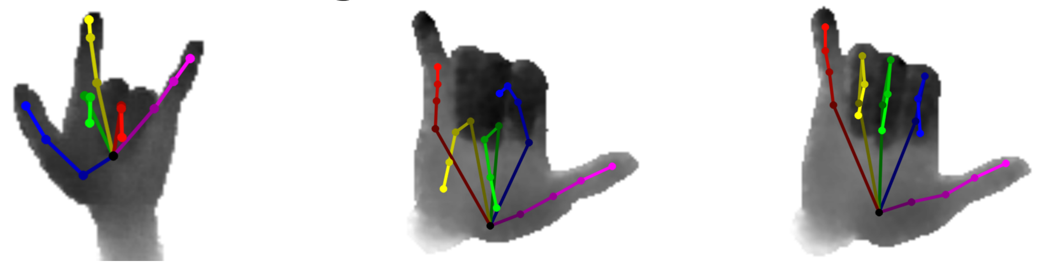
Markus Oberweger, Gernot Riegler, Paul Wohlhart and Vincent Lepetit

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

CVPR 2016

## Motivation

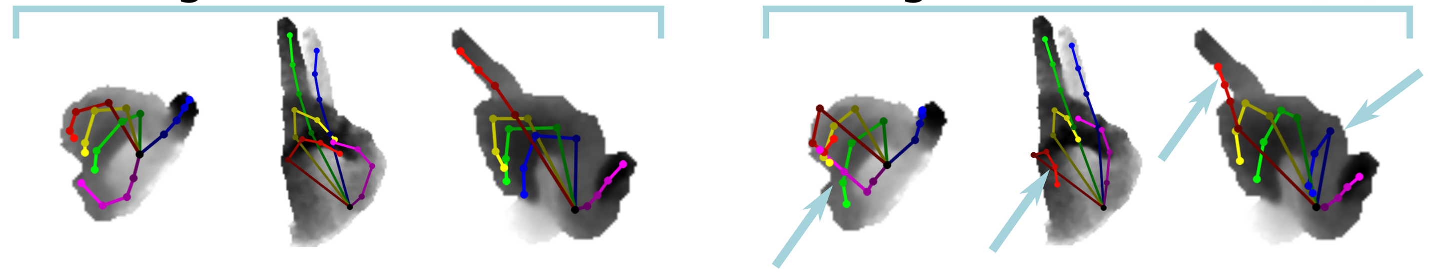
- SOTA hand pose estimation methods are data-driven, but how do we get **accurate** data?



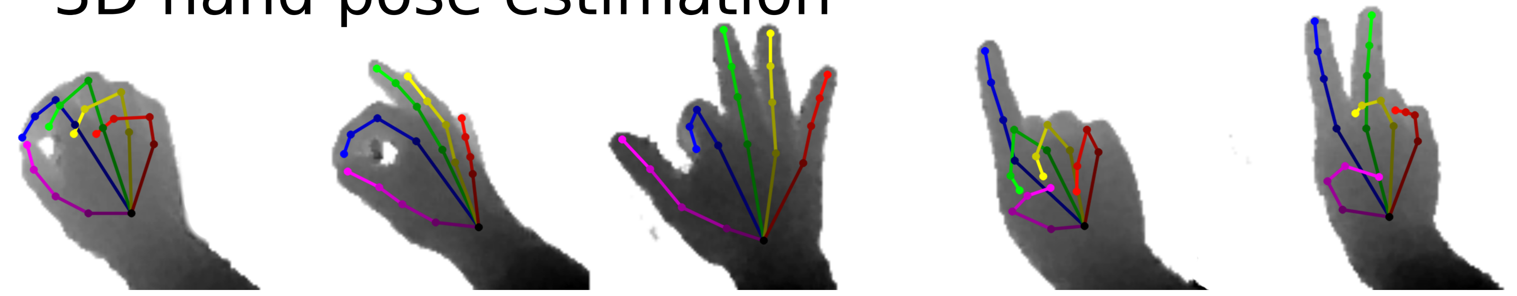
ICVL [3] dataset   MSRA [2] dataset   Our annotation

- Training SOTA pose estimator [1] with *better* annotations

Using our annotations   Using annotations of [2]

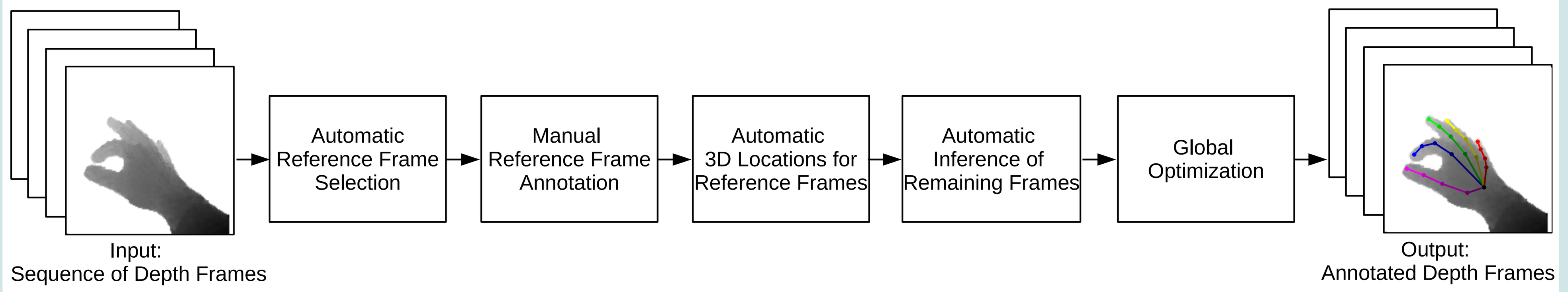


- Goal: Accurate 3D training data for single view depth sequences from sparse 2D annotations
- Reduce time spent on annotations by a factor of 10
- We provide a new dataset for egocentric 3D hand pose estimation



- **Code and dataset are available online**

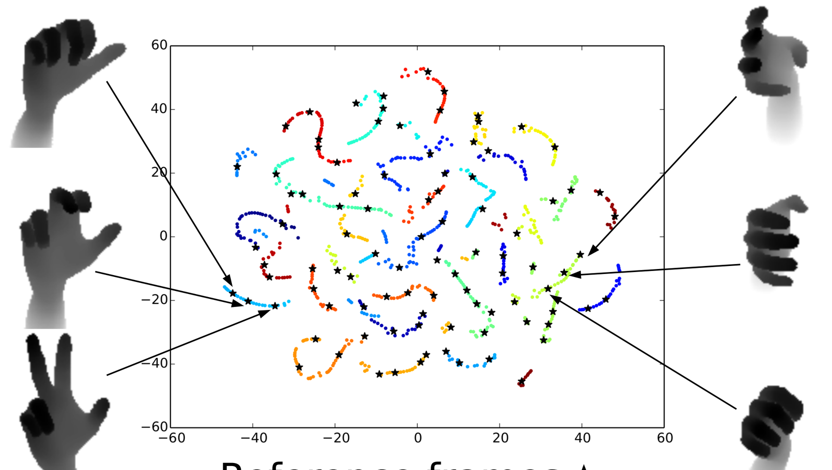
## Method Overview




## 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]

$$\max_{\mathcal{R}} f(\mathcal{R}) \quad \text{s.t.} \quad |\mathcal{R}| < M \qquad f(\mathcal{R}) = |\{i \in [1;N] \; \min_{j\in\mathcal{R}} d(\mathcal{D}_i, \mathcal{D}_j) < \delta\}|$$

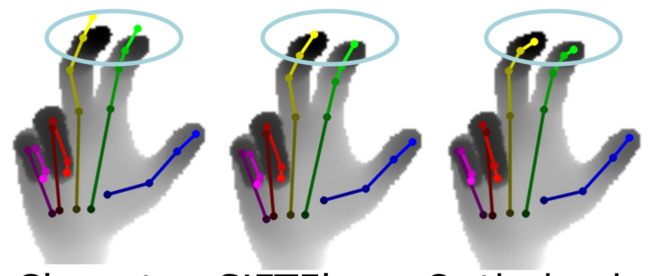


Reference frames★

- 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

$$\begin{aligned} &\arg\min_{\{L_k\}_{k=1}^K} \sum_{k=1}^K vis_k \|\text{proj}(L_k) - l_k\|_2^2 \\ \text{s.t.} \quad & \forall k \;\; \|L_k - L_{p(k)}\|_2^2 = d_{k,p(k)}^2 \\ & \forall k \;\; vis_k = 1 \Rightarrow \mathcal{D}[l_k] < z(L_k) < \mathcal{D}[l_k] + \epsilon \\ & \forall k \;\; vis_k = 1 \Rightarrow (L_k - L_{p(k)})^\top \cdot c_k > 0 \\ & \forall k \;\; vis_k = 0 \Rightarrow z(L_k) > \mathcal{D}[l_k] \end{aligned}$$

- 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]



Closest   SIFTFlow   Optimized

  - Optimize for 3D locations:
    - Maximize similarity of joint appearance in depth map between initialized and not initialized frame
    - Skeleton constrained by bone length

$$\begin{aligned} &\arg\min_{\{L_{\hat{c},k}\}_k} \sum_k \text{dissim}(\mathcal{D}_{\hat{c}}, \text{proj}(L_{\hat{c},k}); \mathcal{D}_{\hat{a}}, l_{\hat{a},k})^2 \\ \text{s.t.} \quad & \forall k \quad \|L_{\hat{c},k} - L_{\hat{c},p(k)}\|_2^2 = d_{k,p(k)}^2 \end{aligned}$$

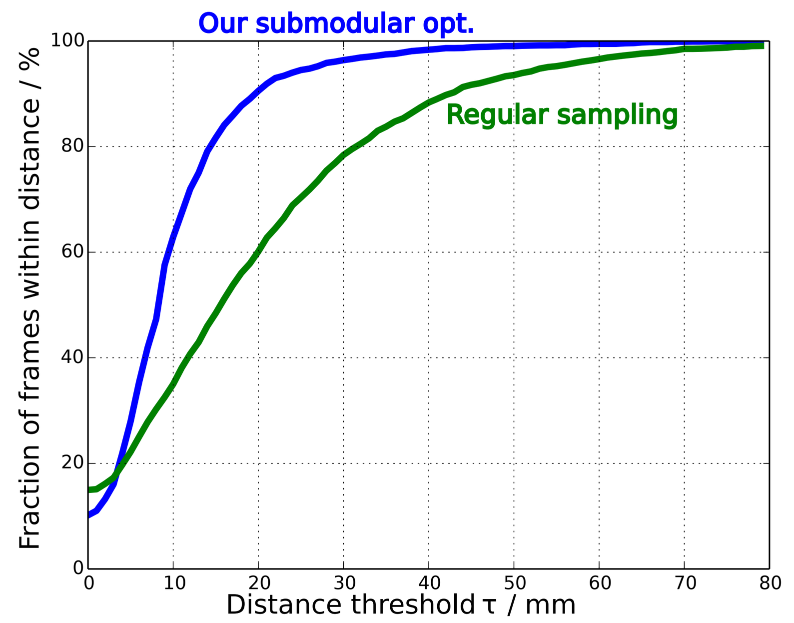
- Global optimization for all 3D locations
  - Maximize similarity of joint appearance in depth map between reference and non-reference frame
  - Enforce temporal smoothness
  - Ensure consistency with 2D user annotations
  - Skeleton constrained by bone length

$$\begin{aligned} &\sum_{i\in[1;N]\setminus\mathcal{R}} \sum_k \text{dissim}(\mathcal{D}_i, \text{proj}(L_{i,k}); \mathcal{D}_{\hat{i}}, l_{\hat{i},k})^2 \; + \\ &\lambda_M \sum_i \sum_k \|L_{i,k} - L_{i+1,k}\|_2^2 \; + \\ &\lambda_P \sum_{r\in\mathcal{R}} \sum_k vis_{r,k} \|\text{proj}(L_{r,k}) - l_{r,k}\|_2^2 \\ \text{s.t.} \quad & \forall i,k \quad \|L_{i,k} - L_{i,p(k)}\|_2^2 = d_{k,p(k)}^2 \end{aligned}$$

More details can be found in the paper.

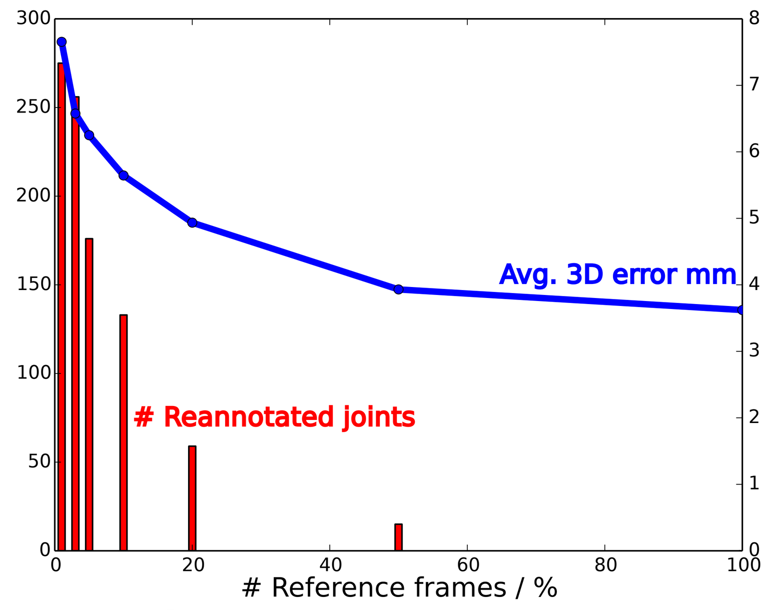
## Results

- Efficient reference frame selection




Our selection method:
- Frames closer to reference in 3D → easier propagation to other frames
- Close to exact solution
- Fast runtime
- Scales well to large datasets

- Accuracy vs. annotation effort
  - Best results when annotating all
  - Only minor decrease in accuracy for a fraction of annotations




- Accurate 3D locations for reference frames
  - Intuitive and fast to annotate (45s per frame)
  - Robust to annotation noise

| Used annotations | Visible joints: Avg. / median | All joints: Avg. / median |
|---|---|---|
| 2D locations | 12.86 / 8.96mm | 19.98 / 13.29mm |
| 2D & visibility | 3.94 / 3.18mm | 6.20 / 3.41mm |
| 2D & vis & z-order | **2.97 / 2.93mm** | **3.65 / 2.98mm** |
| All + 2D noise | 3.70±0.71 / 2.59±0.21mm | 4.29±0.63 / 2.56±0.23mm |

- Accuracy of full pipeline for 10% reference frames
  - 3D locations of closest reference are not accurate enough
  - Gain of SIFTFlow small, but useful for joint appearance term

| Method | Avg. / median error |
|---|---|
| Closest reference | 11.50 / 5.58mm |
| Aligned with SIFTFlow | 11.40 / 5.40mm |
| Frame optimization | 5.76 / 4.34mm |
| Global optimization | **5.53 / 4.23mm** |

[1] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands Deep in Deep Learning for Hand Pose Estimation. In Proc. of CVWW, 2015.
[2] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded Hand Pose Regression. In CVPR, 2015.
[3] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture. In CVPR, 2014.
[4] C. Liu, J. Yuen, and A. Torralba. SIFT Flow: Dense Correspondence Across Scenes and Its Applications. PAMI, 33(5), 2011.
[5] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An Analysis of Approximations for Maximizing Submodular Set Functions - I. Mathematical Programming, 14(1), 1978.