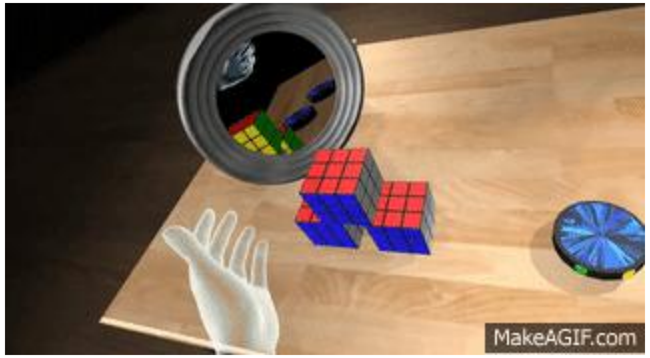
Occlusion aware hand pose recovery from sequences of depth images

Meysam Madadi 1,2 , Sergio Escalera 1,3 , Alex Carruesco 4 ,
Carlos Andujar 4 , Xavier Baró 1,5 , Jordi Gonzàlez 1,2

1 Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra (Barcelona), Catalonia Spain
2 Dept. of Computer Science, Univ. Aut`onoma de Barcelona (UAB), 08193 Bellaterra, Catalonia Spain
3 Depl. Mathematics and Informatics, Universitat de Barcelona, Catalonia, Spain
4 ViRVIG-Moving Research Group, UPC-BarcelonaTech
5 Universitat Oberta de Catalunya, Catalonia, Spain

Introduction

- What is hand pose recovery?
- And applications:
  - Human-computer interaction,
  - Virtual reality,
  - Robot learning.

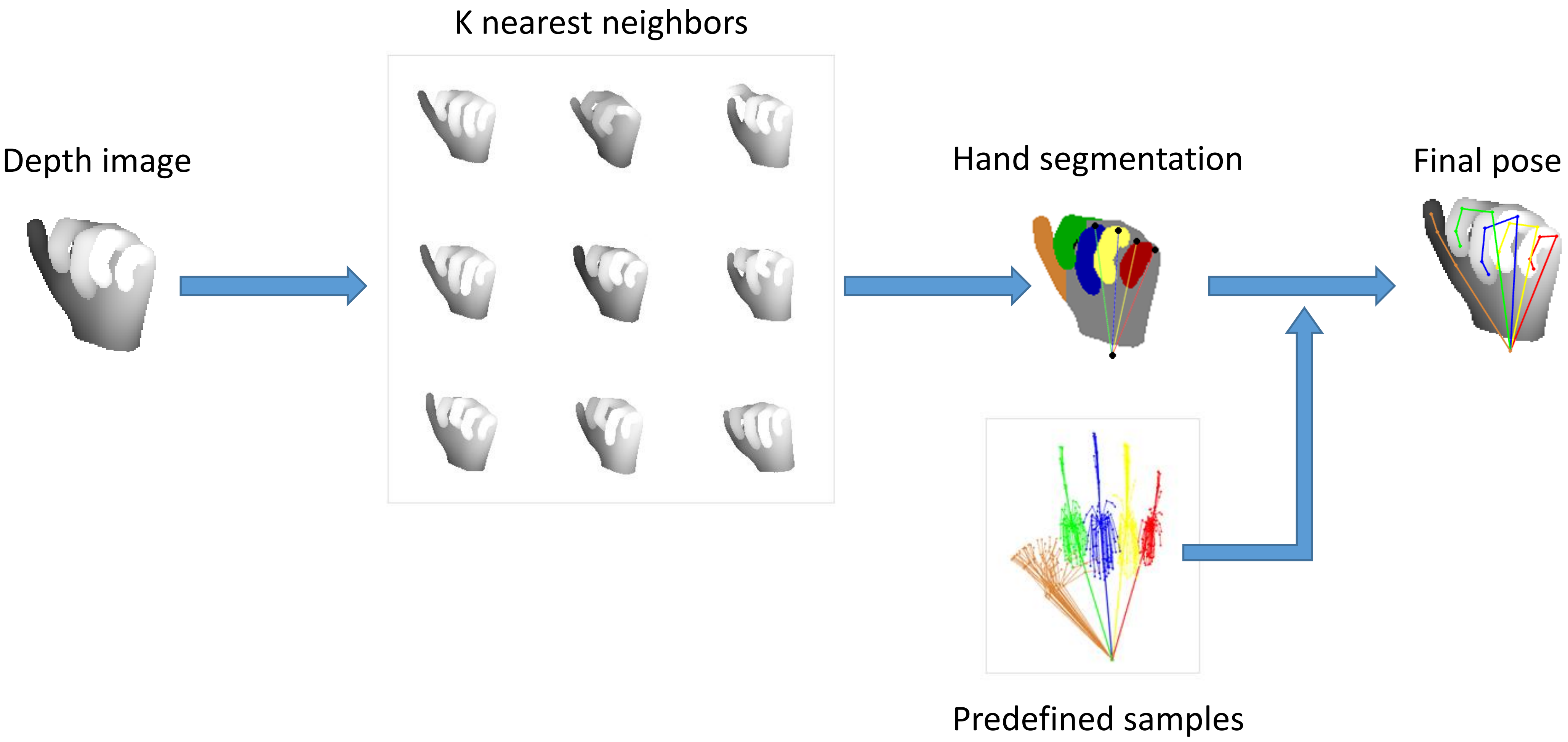


Outline

- Single frame pose recovery
- Temporal pose recovery
- Results

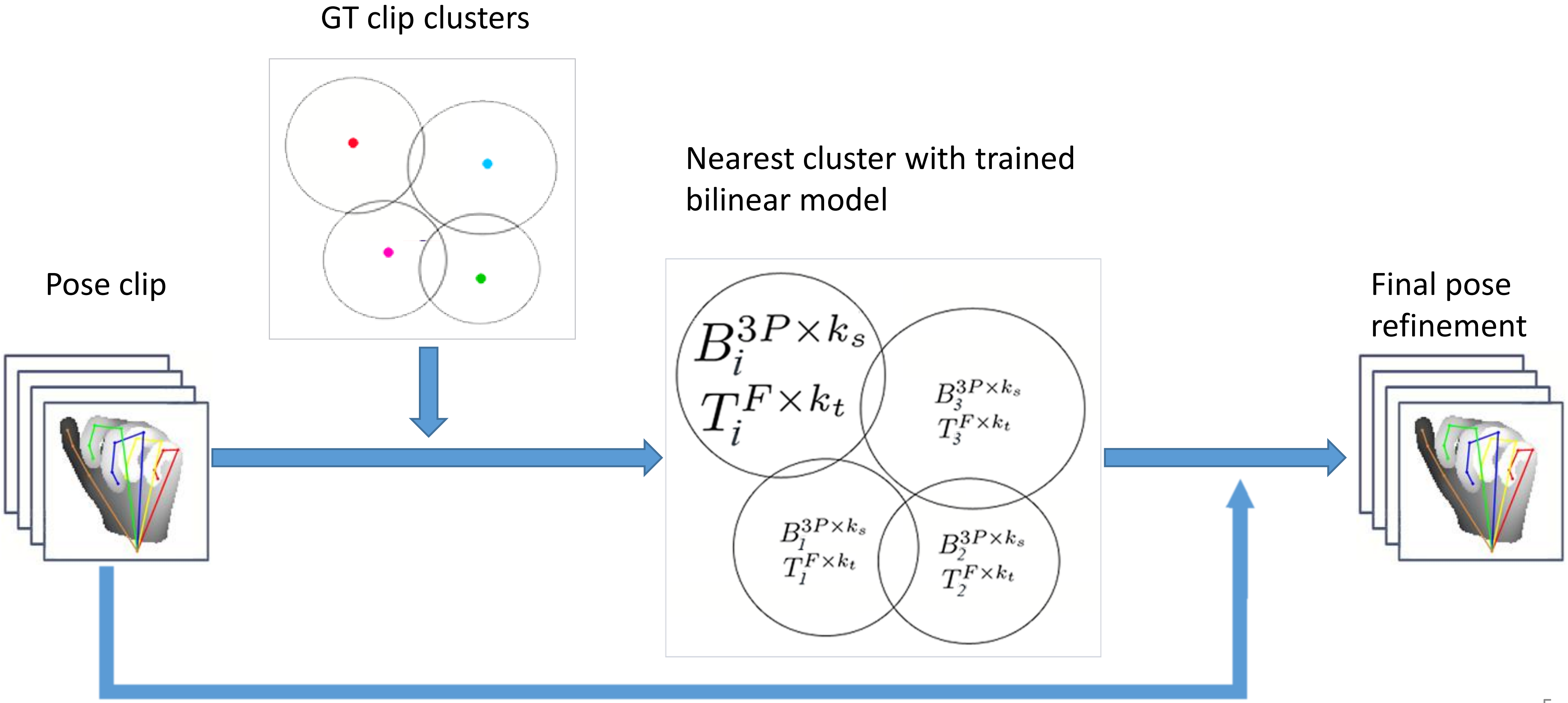
System overview and pipeline

- Single frame hand pose recovery



System overview and pipeline

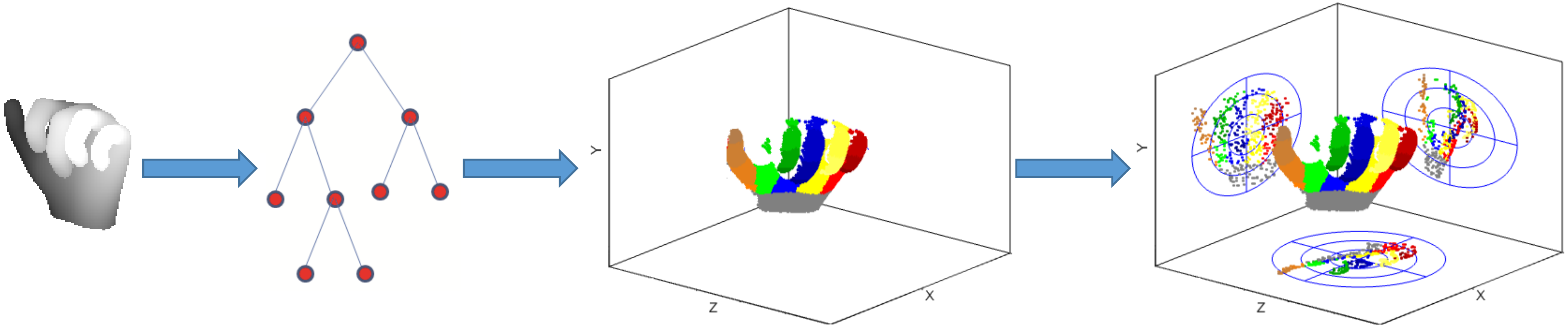
- Spatial-temporal hand pose recovery



Single frame hand pose recovery

- K nearest neighbors are extracted based on a novel descriptor.

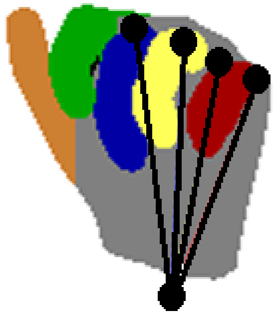
Depth image | Random forest | Class probabilities | Shape context of each plane over class probabilities



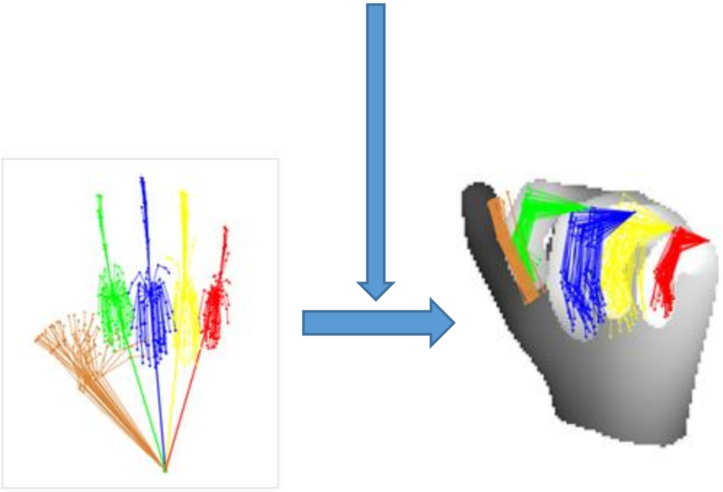
$$H_{xy}(k,c) = \sum_{i=1}^{N} \{R_{ic} | (P_i^{xy} - q^{xy}) \in bin_{xy}(k)\},$$

Single frame hand pose recovery

- K nearest neighbors are aligned to hand point cloud to:
  1. Segment hand into palm and fingers,
  2. Extract palm joints.

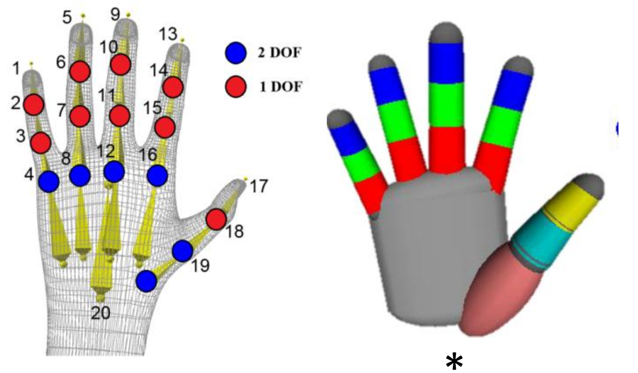


- A set of candidate fingers are selected given:
  1. Hand segments and palm joints,
  2. A predefined set of sample fingers,
  3. A set of simple rules:
     - Joints must not be located outside the hand mask,
     - A joint must not have a depth lower than the hand surface.



Single frame hand pose recovery

- We fit a finger model on hand depth image for each finger given:
  - Hand segments,
  - Selected finger candidates,
  - A discrepancy function ***E***:



*

$$E(h, I) = w_1 E_1 + w_2 E_2 + w_3 E_3$$

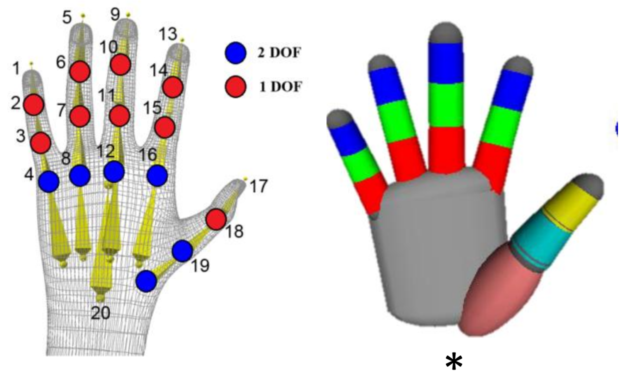
1 - Normalized overlapping area between finger model and finger segment

* I.Oikonomidis, N.Kyriazis, and A.A.Argyros. Markerless and efficient 26-dof hand pose recovery. In ACCV, 2010

Single frame hand pose recovery

- We fit a finger model on hand depth image for each finger given:
  - Hand segments,
  - Selected finger candidates,
  - A discrepancy function ***E***:

$$E(h, I) = w_1 E_1 + w_2 E_2 + w_3 E_3$$



*

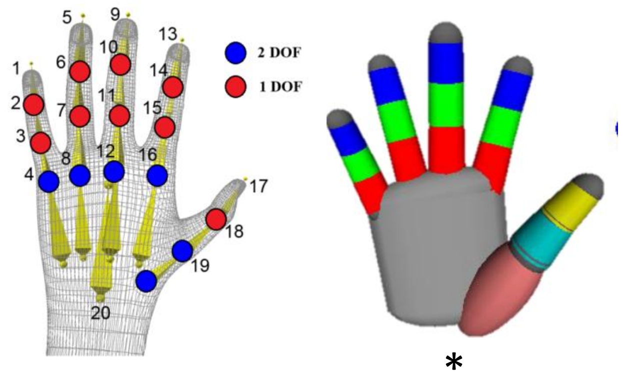
Normalized depth discrepancy between finger model and finger segment

* I.Oikonomidis, N.Kyriazis, and A.A.Argyros. Markerless and efficient 26-dof hand pose recovery. In ACCV, 2010

Single frame hand pose recovery

- We fit a finger model on hand depth image for each finger given:
  - Hand segments,
  - Selected finger candidates,
  - A discrepancy function ***E***:

$$E(h, I) = w_1 E_1 + w_2 E_2 + w_3 E_3$$



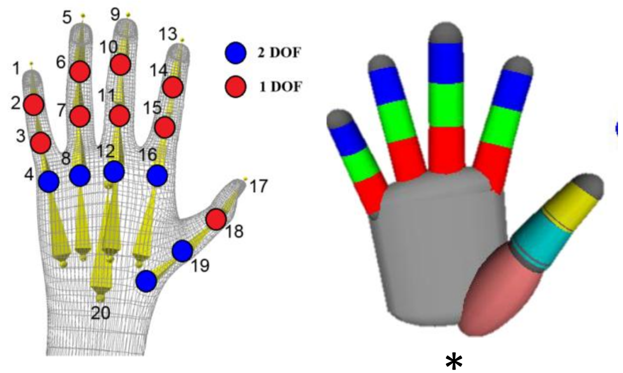
*

Penalizing finger model collision with background fingers

* I.Oikonomidis, N.Kyriazis, and A.A.Argyros. Markerless and efficient 26-dof hand pose recovery. In ACCV, 2010

Single frame hand pose recovery

- We fit a finger model on hand depth image for each finger given:
  - Hand segments,
  - Selected finger candidates,
  - A discrepancy function ***E***



*

•Greedy approach: each finger candidate is applied of *E*, and the one with minimum value is selected as prediction
•PSO: particles are initialized with finger parameters *h*, a new prediction *h** is found after optimization

* I.Oikonomidis, N.Kyriazis, and A.A.Argyros. Markerless and efficient 26-dof hand pose recovery. In ACCV, 2010

Spatial-temporal hand pose recovery

- To incorporate temporal data, we concatenate estimated poses from last ***F*** frames into clip matrix $Q \in \mathbb{R}^{F \times 5D}$
- ***Q*** can be factorized through $Q = TCB^T$ *

Learned trajectory bases using discrete cosine transform

* I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear spatiotemporal basis models. TOG, 31(17), 2012.

Spatial-temporal hand pose recovery

- To incorporate temporal data, we concatenate estimated poses from last ***F*** frames into clip matrix $Q \in \mathbb{R}^{F \times 5D}$
- ***Q*** can be factorized through $Q = TCB^T$ *

Learned shape bases using **SVD**

* I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear spatiotemporal basis models. TOG, 31(17), 2012.

Spatial-temporal hand pose recovery

- To incorporate temporal data, we concatenate estimated poses from last ***F*** frames into clip matrix $Q \in \mathbb{R}^{F \times 5D}$
- ***Q*** can be factorized through $Q = TCB^T$ *

Coefficient matrix

* I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear spatiotemporal basis models. TOG, 31(17), 2012.

Spatial-temporal hand pose recovery

- To incorporate temporal data, we concatenate estimated poses from last ***F*** frames into clip matrix $Q \in \mathbb{R}^{F\times 5D}$
- ***Q*** can be factorized through $Q = TCB^T$ *
- The goal is to minimize an objective function over coefficient matrix ***C.***

→ **Problem:** Linear models like **PCA** and **SVD** are sensitive to the **distribution of data**.

→ **Solution: Clusterize** clips into smaller and more inter-correlated categories,

→ **and approximate best cluster** over extracted **K** nearest clusters.

* I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear spatiotemporal basis models. TOG, 31(17), 2012.

Spatial-temporal hand pose recovery

- We define objective function as:

$$\underset{C}{argmin} \underbrace{\sum_{f=1}^{F}\sum_{i=1}^{5D} V_{fi}\left|Q_{fi} - [TCB^T]_{fi}\right|}_{\text{Reconstruction error with respect to visible joints}} + \beta \sum_{f=1}^{F-1} \Psi^{f,f+1}$$

Spatial-temporal hand pose recovery

- We define objective function as:

$$\underset{C}{argmin} \sum_{f=1}^{F} \sum_{i=1}^{5D} V_{fi} \left| Q_{fi} - \left[TCB^T\right]_{fi} \right| + \beta \sum_{f=1}^{F-1} \Psi^{f,f+1}$$

Smoothness function of consequence frames

Spatial-temporal hand pose recovery

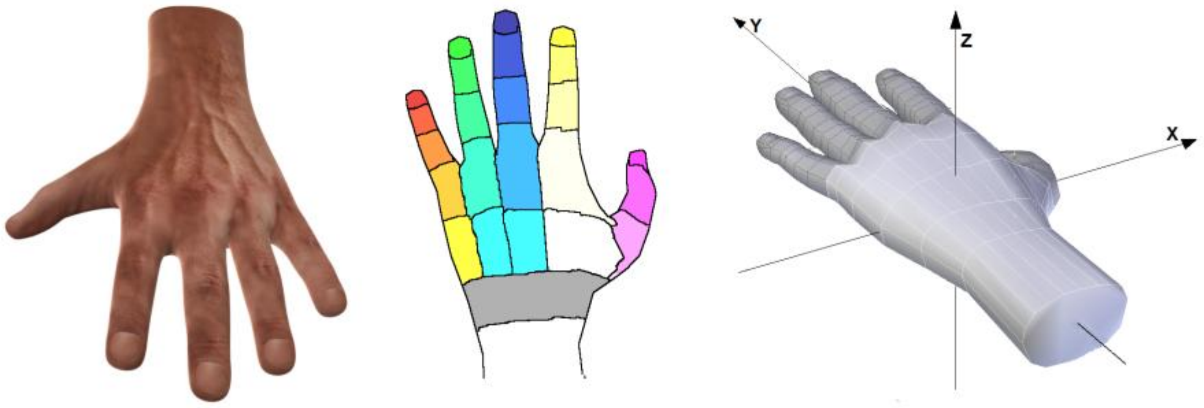
- We define objective function as:

$$\underset{C}{argmin} \sum_{f=1}^{F} \sum_{i=1}^{5D} V_{fi} \left| Q_{fi} - [TCB^T]_{fi} \right| + \beta \sum_{f=1}^{F-1} \Psi^{f,f+1}$$

- Particle swarm optimization is applied to minimize objective function, initial particles are defined as random guesses near to ***Q***.

Dataset

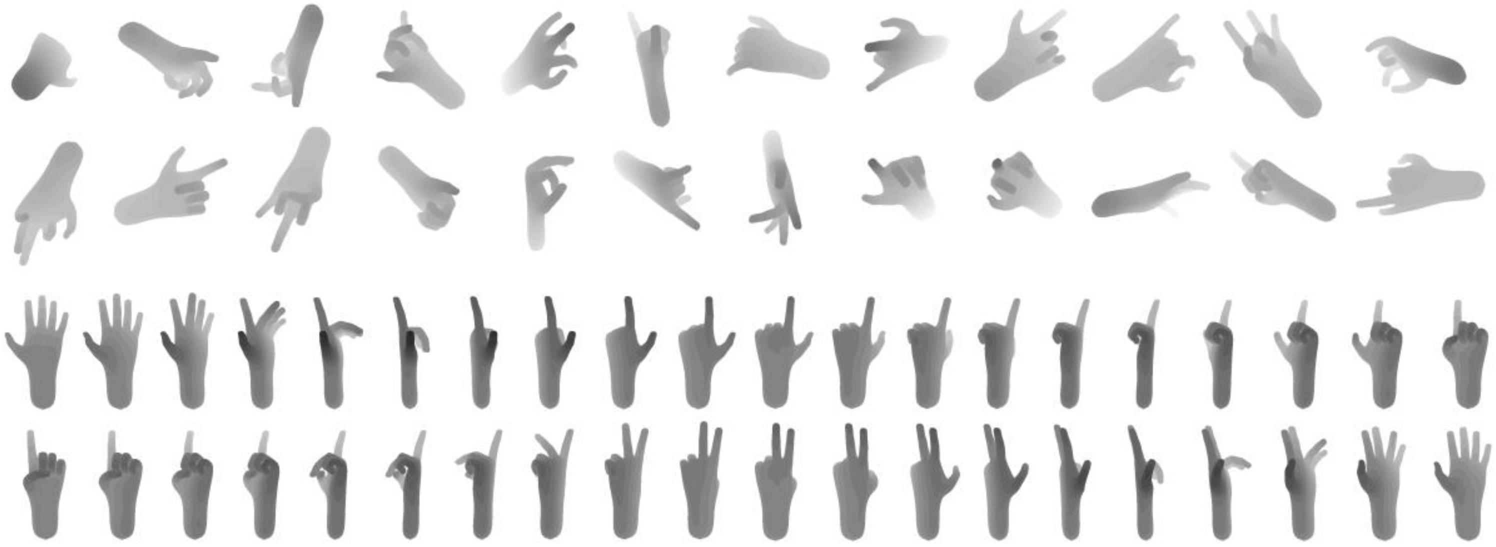
- Current datasets are mainly designed for:
  - Front view hand pose recovery,
  - Single frame hand pose recovery.
- We generated a synthetic hand dataset with natural finger movements and high degree of occlusion, consisting of +1M frames.



http://chalearnlap.cvc.uab.es/dataset/25/description/

Dataset

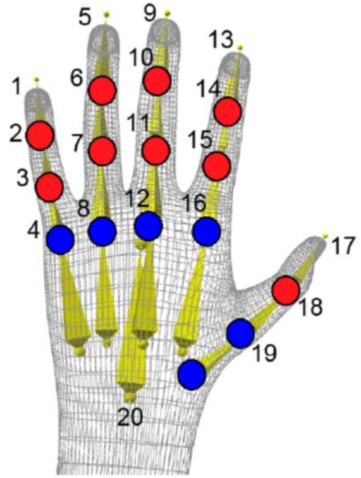
- Current datasets are mainly designed for:
  - Front view hand pose recovery,
  - Single frame hand pose recovery.
- We generated a synthetic hand dataset with natural finger movements and high degree of occlusion, consisting of +1M frames.

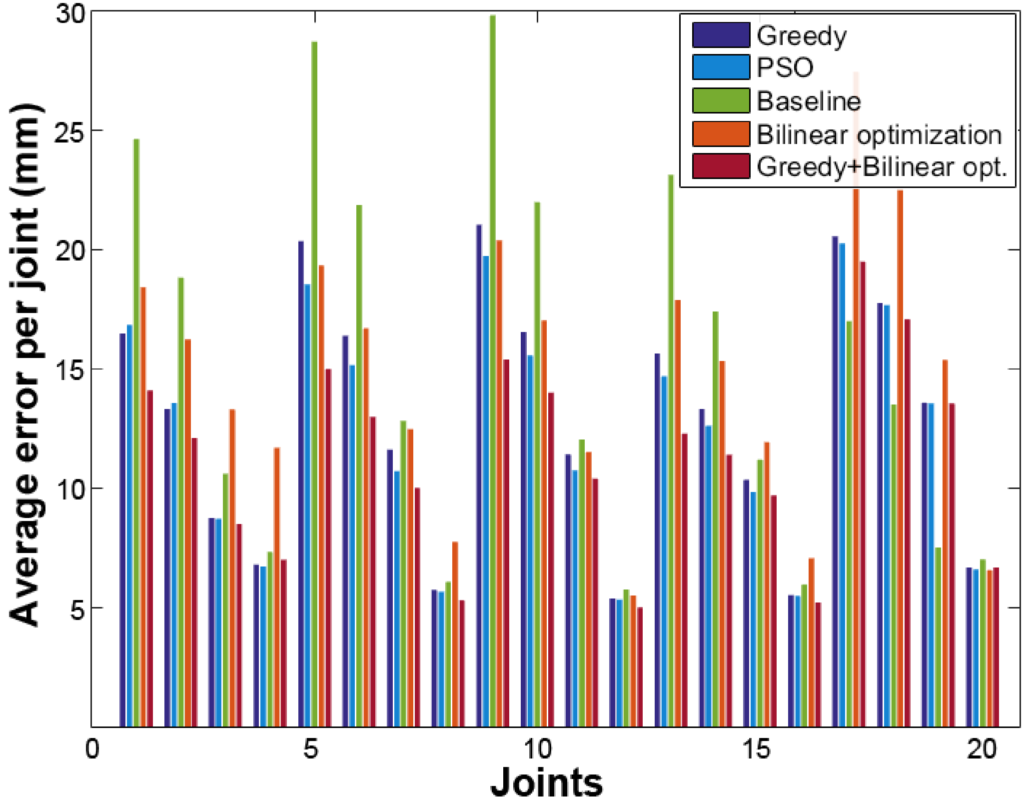


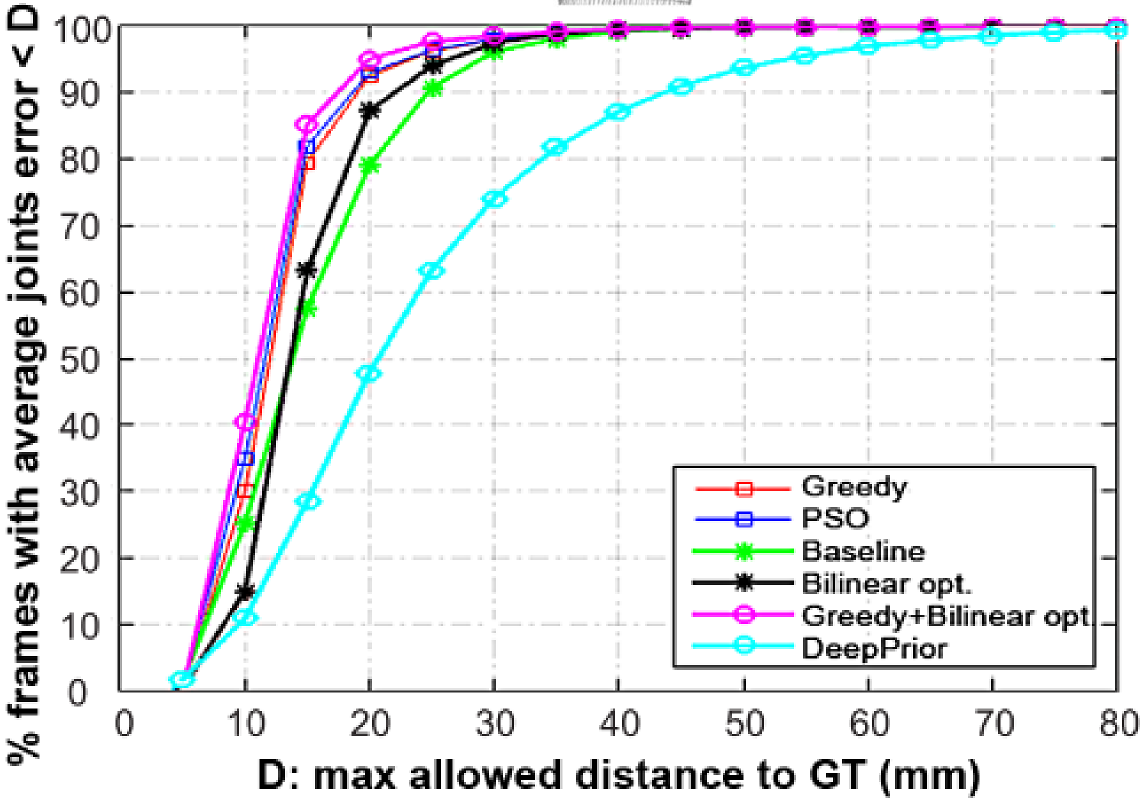
http://chalearnlap.cvc.uab.es/dataset/25/description/

Results

- Quantitative results on our dataset (baseline is a 1NN approach)



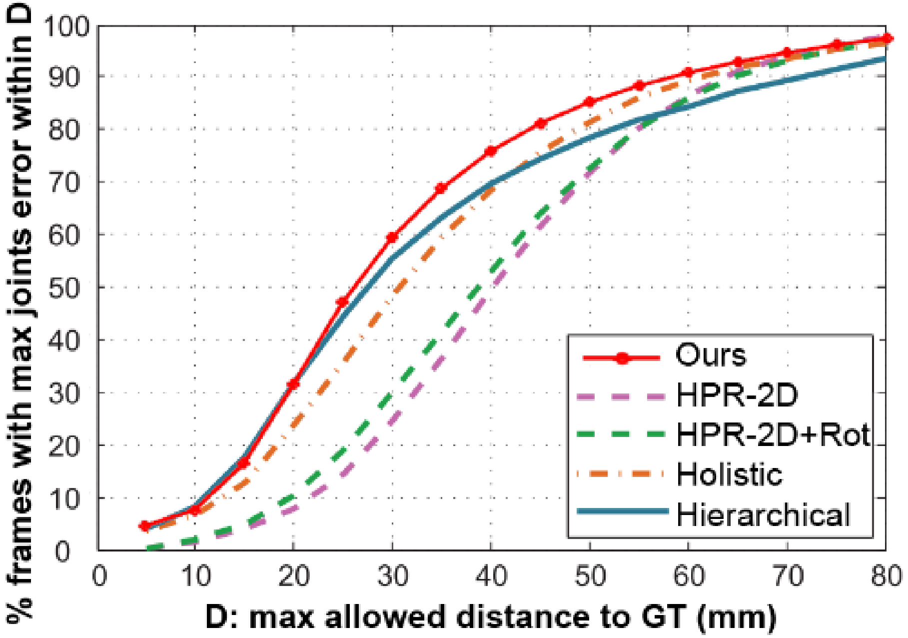




M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. CVWW, 2015.

Results

- Quantitative results on MSRA* dataset



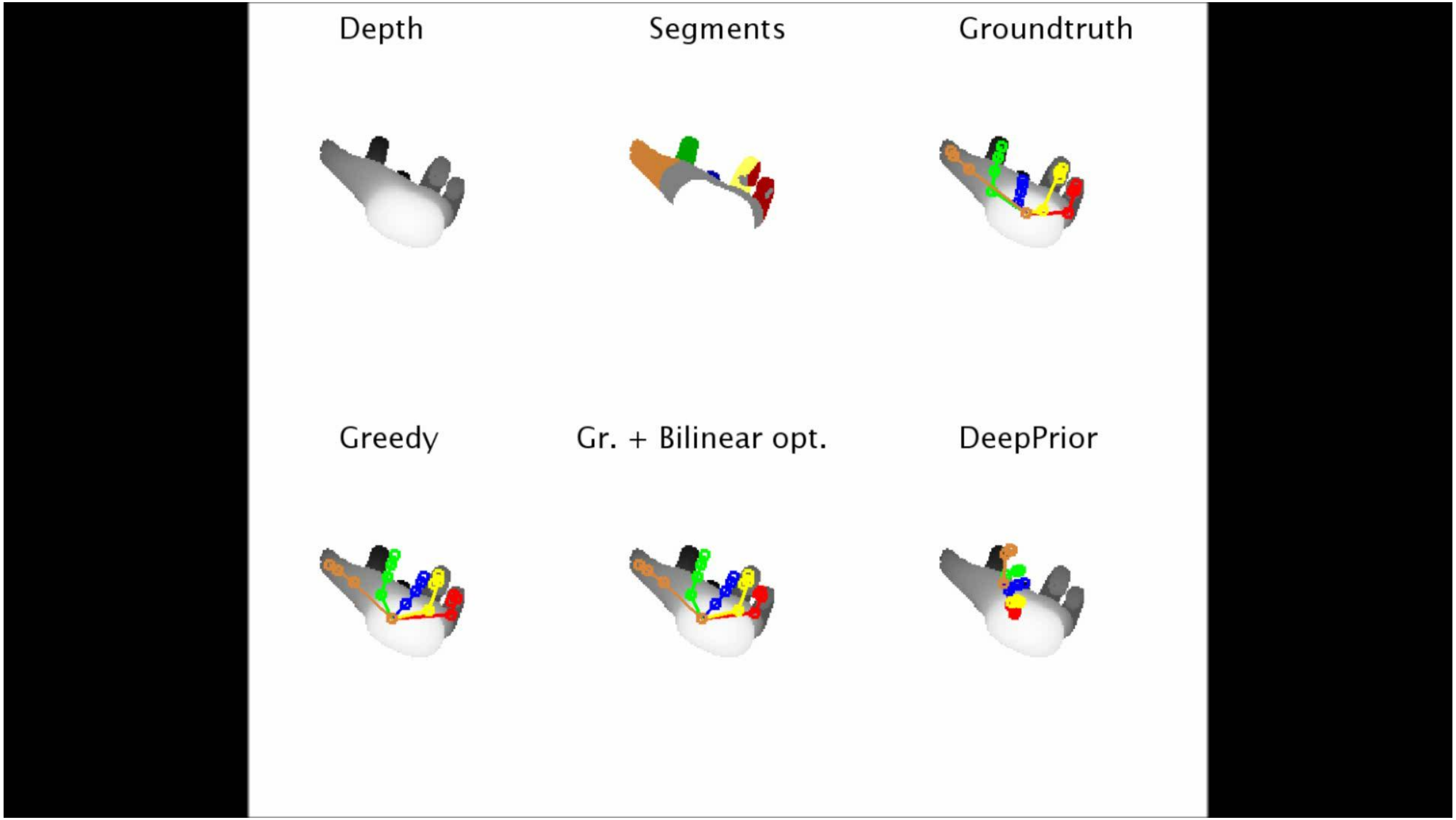
X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In CVPR, 2015.

	IndexR	IndexT	MiddleR	MiddleT	RingR	RingT	LittleR	LittleT	ThumbT	Mean
Oikonomidis et al. [14]	31.0	56.0	32.9	56.0	32.9	49.3	35.1	53.7	22.2	38.2
Choi et al. [3]	22.6	43.5	24.0	44.9	23.1	43.1	21.8	39.5	31.1	29.8
Ge et al. [12]	11.5	16.0	9.0	15.6	9.9	15.1	13.2	16.0	16.7	13.0
Ours (KNN+ICP)	9.5	17.3	7.7	17.1	8.3	15.5	10.6	17.7	14.8	12.8

[3] C. Choi, A. Sinha, J. H. Choi, S. Jang, and K. Ramani. A collaborative filtering approach to real-time hand pose estimation. ICCV, 2015.
[12] J. Y. Liuhao Ge, Hui Liang and D. Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. CVPR, 2016.
[14]I. Oikonomidis, N. Kyriazis, and A. Argyros. Efcient model-based 3d tracking of hand articulations using kinect. BMVC, pages 101.1–101.11, 2011.

Results

- Qualitative results on our dataset



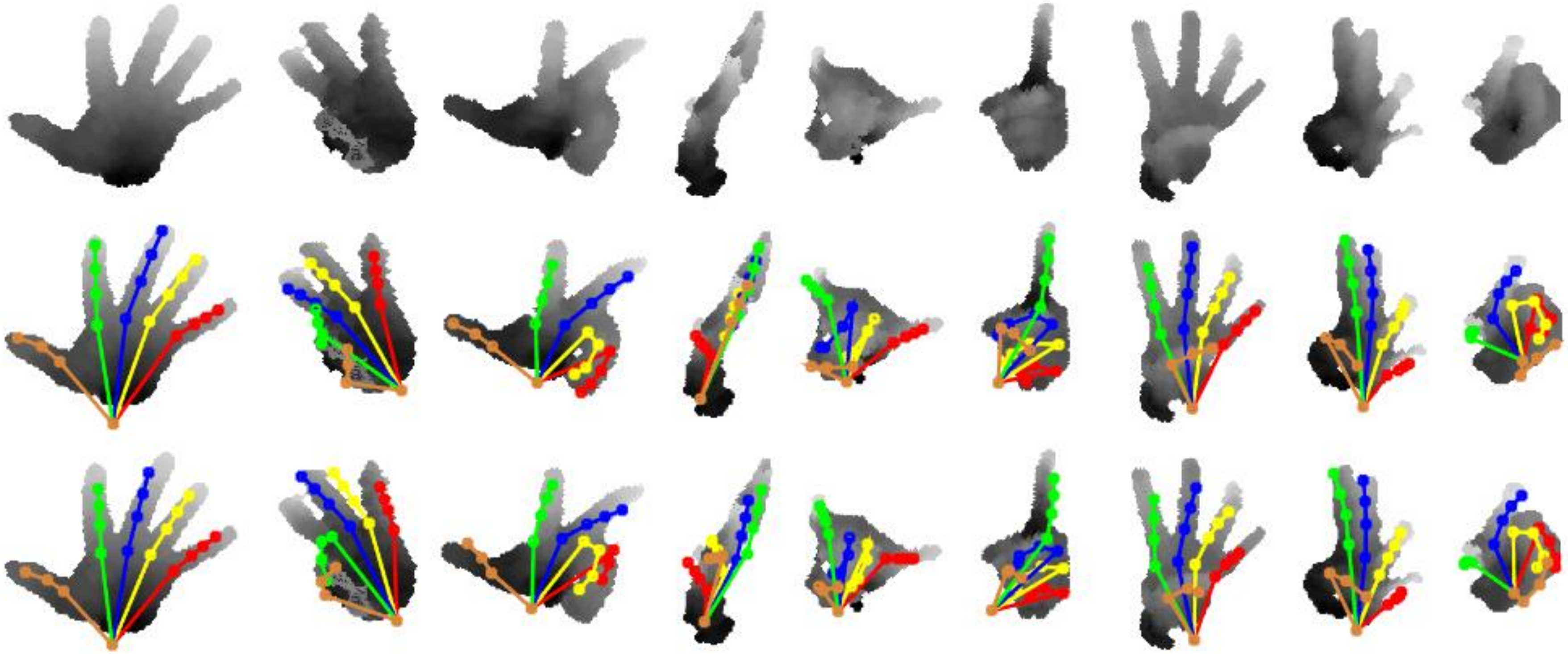
Results

- Qualitative results on MSRA dataset

Depth

GT

Predicted

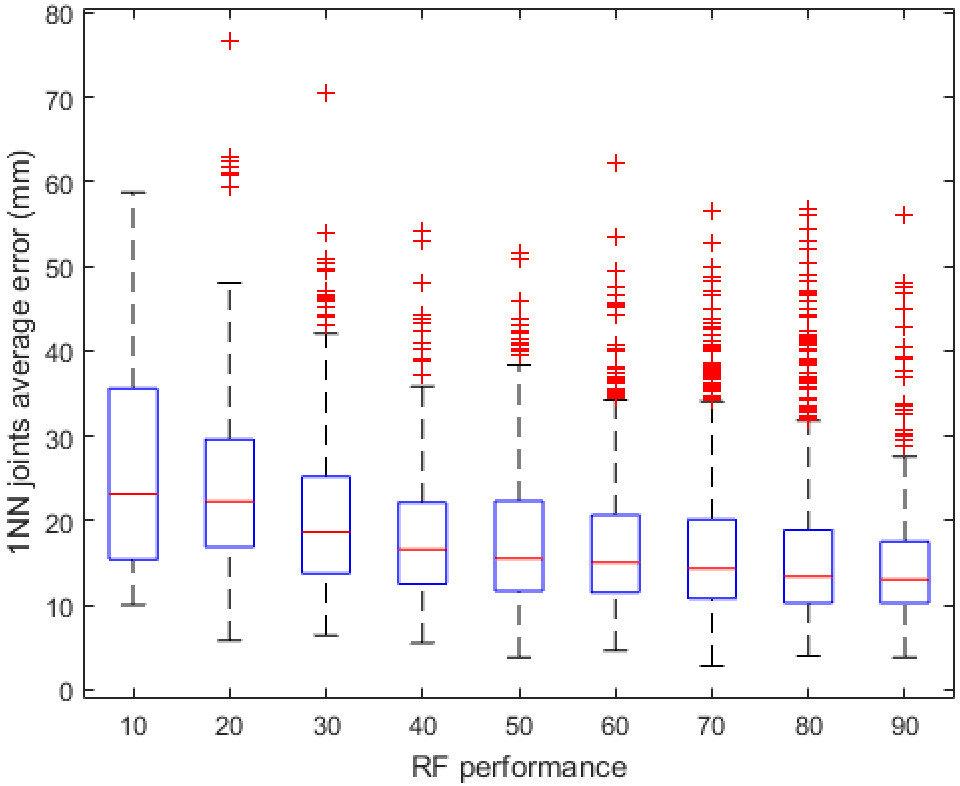


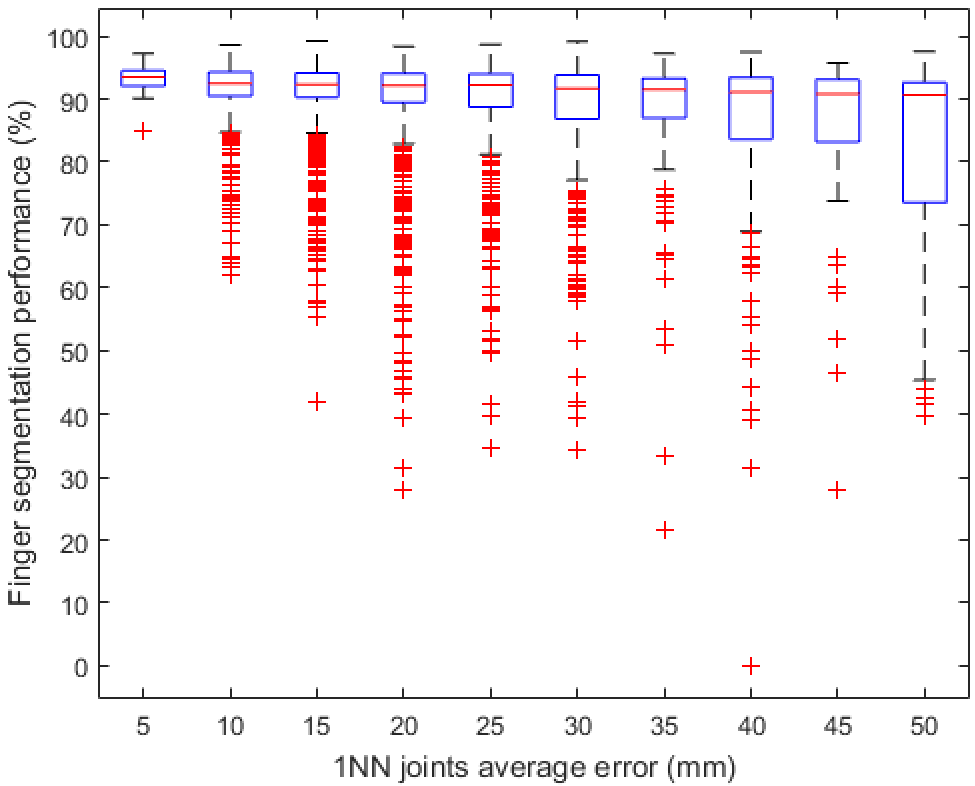
Results

- Components analysis

Joint estimation robust against RF performance

Even if nearest neighbors are non-accurate, after ICP, fingers segmentation are accurate

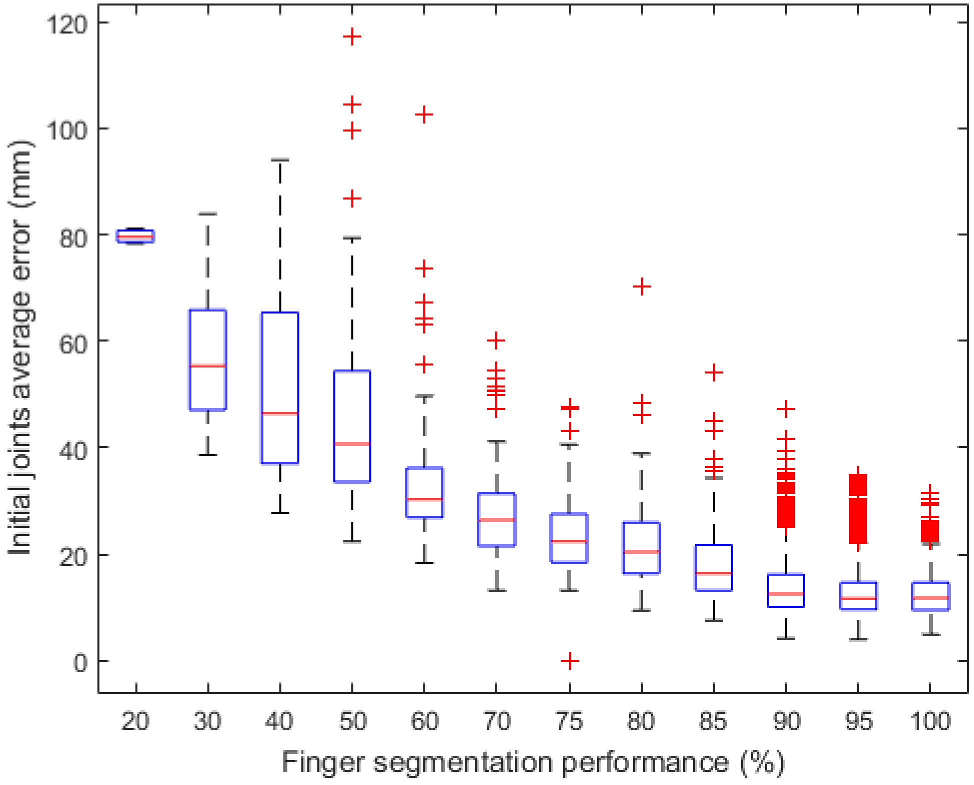




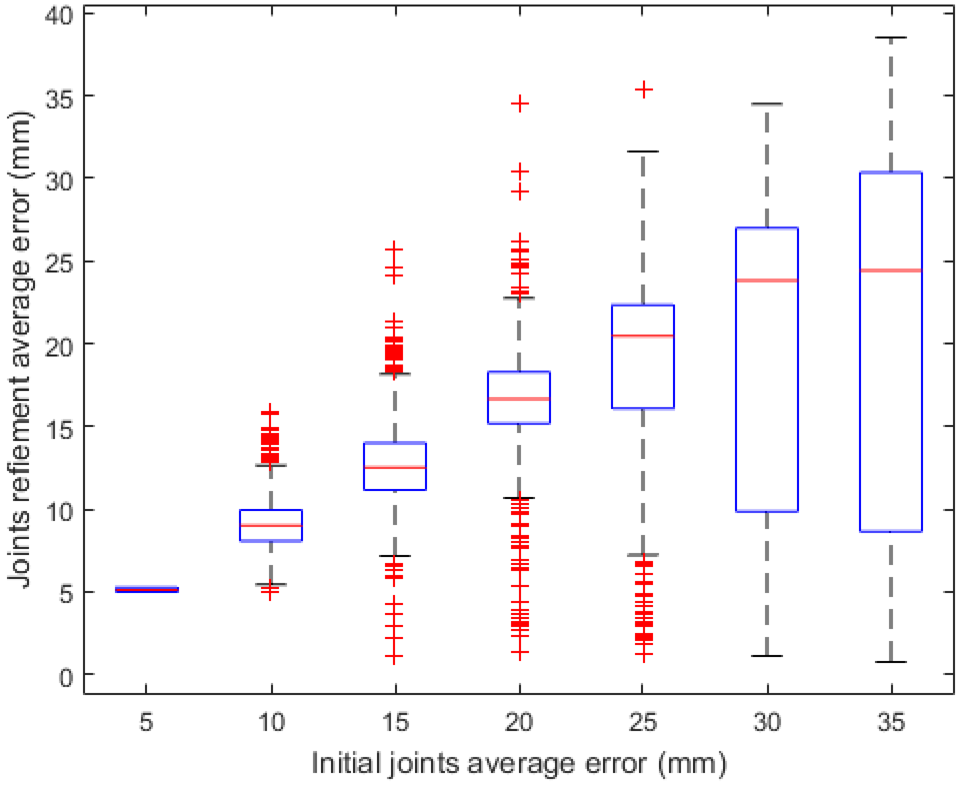
Results

- Components analysis

Joint estimation relation to initial finger segmentation



Joint temporal refinement based on initial static pose error



Conclusions

- We created a synthetic hand dataset with huge variabilities of pose and viewpoint (http://chalearnlap.cvc.uab.es/dataset/25/description/),
- We created a 2.5D shape descriptor,
- By applying nearest neighbors, we efficiently:
  - segmented hand,
  - extracted palm joints,
  - and then sampled a number of candidates.
- We fitted finger models on the hand in a **single frame** including spatial optimization constraints,
- We refined joint estimates, including occluded joints, using **spatio-temporal** linear models,
- Our model is capable of recovering pose in different viewpoints and pose.

Thank you for you attention!