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. Autònoma 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

Depth image ➔ K nearest neighbors ➔ Hand segmentation ➔ Final pose ➔ Predefined samples
System overview and pipeline

- Spatial-temporal hand pose recovery

GT clip clusters

Nearest cluster with trained bilinear model

Pose clip

Final pose refinement
Single frame hand pose recovery

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

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

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

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.
Spatial-temporal hand pose recovery

• We define objective function as:

\[
\arg\min_C \sum_{f=1}^F \sum_{i=1}^{5D} V_{f_i} |Q_{f_i} - [T C B^T]_{f_i}| + \beta \sum_{f=1}^{F-1} \Psi_{f,f+1}
\]

Reconstruction error with respect to visible joints
Spatial-temporal hand pose recovery

• We define objective function as:

\[
\text{argmin}_C \sum_{f=1}^{F} \sum_{i=1}^{5D} V_{f,i} \left| Q_{f,i} - [TCB^T]_{f,i} \right| + \beta \sum_{f=1}^{F-1} \Psi_{f,f+1}
\]

Smoothness function of consequence frames
Spatial-temporal hand pose recovery

• We define objective function as:

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

• 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)

Results

• Quantitative results on MSRA* dataset

% frames with max joints error within D

D: max allowed distance to GT (mm)


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 your attention!