Efficient Model-based Tracking of the Articulated Motion of Hands

Iason Oikonomidis, Nikolaos Kyriazis, Antonis A. Argyros

Institute of Computer Science, FORTH, Greece

AND

Department of Computer Science, University of Crete, Greece

PROBLEM STATEMENT

Track the 3D position, orientation and full articulation (26 DoFs) of a human hand that possibly manipulates an object, given a sequence of either multi-view or RGB-D frames of the scene.

MOTIVATION

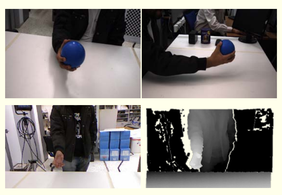
The markerless tracking of hand articulations is a challenging problem with diverse applications such as H.C.I., understanding human grasping, robot learning by demonstration, etc.

MAIN IDEA

Jointly consider the observed scene: extract full-image features and produce full hypotheses about it. Compare hypotheses and observed features in parallel [4]. Use these scores to drive an iterative optimization process using Particle Swarm Optimization (PSO) [5].

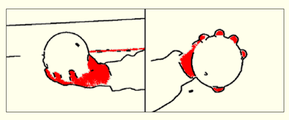
PROPOSED METHOD

Input



- Frames are acquired by a **multi-camera** setup or a **Kinect**.

- **Edge** (black) and **skin color** (red) cues are extracted for the multi-view case.

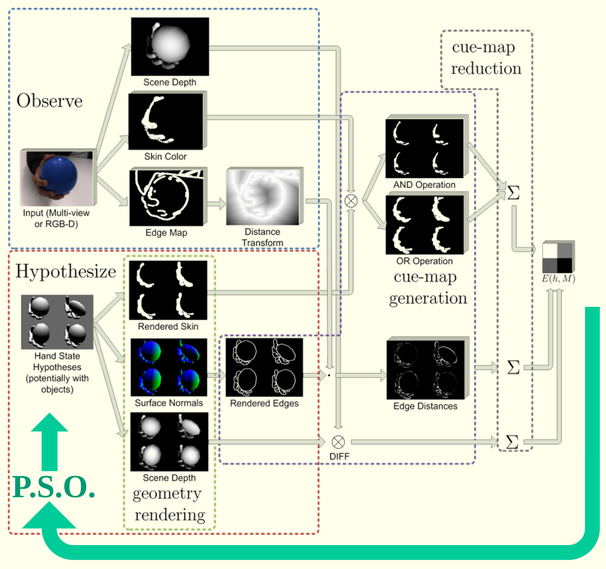


- For the case of Kinect, skin color and depth cues, along with the **temporal continuity** assumption are used to **segment** the hand.

Fit model to data

- Employ a parametric hand model [1].
  - ❑ Comprised of 15 cylinders and 22 spheres.
  - ❑ 26 DoFs: 6D global pose, 20 kinematics angles.

  \+ For the hand-object case, add a parametric object (9 or 10 DoFs).

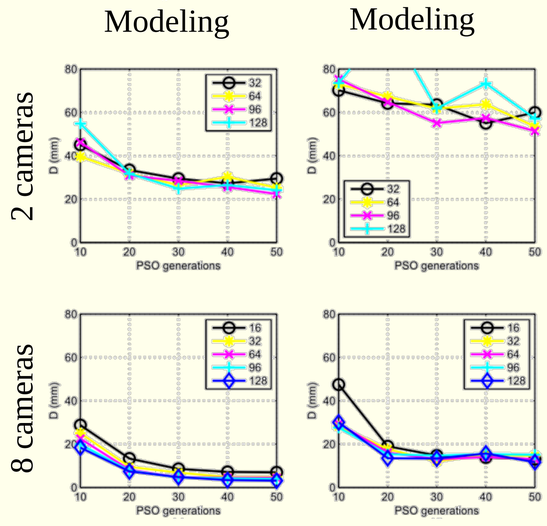


- From a full configuration (all 26 DoFs of the hand model plus potentially the object DoFs), a **skin occupancy map,** an **edge map** and a **depth map** can be **synthesized** by means of rendering.
- These **maps** are used to quantify the **discrepancy** between **observation** and **hypothesis** (objective function).
- The objective function also **penalizes physically implausible** configurations (hand-hand and hand-object collision checking).
- A **variant** of the **PSO** method [5] searches in the model parameter space for the best scoring configuration.
  - ❑ **Efficient** evaluation of **multiple** hypotheses on the **GPU** [4].
- Candidate poses for the **next frame** are obtained by **perturbing** the solution of the **previous frame**.

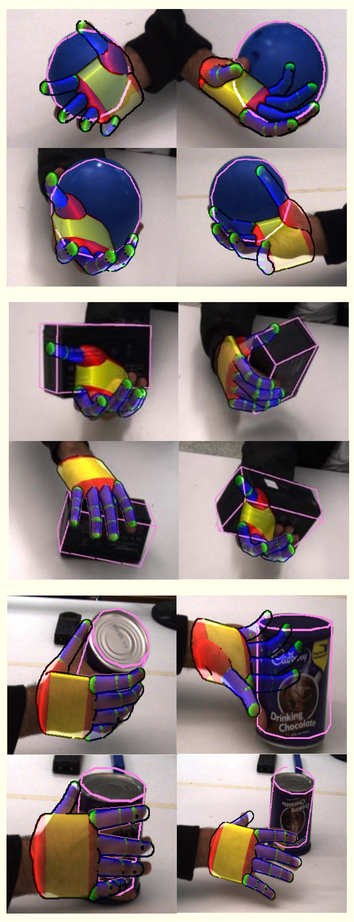
EXPERIMENTAL RESULTS

Input from a Multi-view System

Quantitative evaluation on synthetic data



64 particles and 40 generations for 4 views yield 2fps on a modern PC



Input from Kinect

Single-view depth image
Hand in isolation

64 particles and 30 generations yield 15fps on a modern PC

STRENGTHS OF THE APPROACH

- Occlusions serve as visual cues through modeling.
- Joint optimization: no simplifying assumptions over the problem structure, simultaneous consideration of all parameters.
- Careful design and exploitation of parallelism in a GPU implementation [4] lead to a computationally efficient system that accepts input of multiple modalities [1-3].
- Minimally invasive markerless approach.

KEY REFERENCES

1. Oikonomidis, I., Kyriazis, N., Argyros, A. A. “Markerless and Efficient 26-DOF Hand Pose Recovery”. *ACCV*, 2010.
2. Oikonomidis, I., Kyriazis, N., Argyros, A. A. “Full DOF Tracking of a Hand Interacting with an Object by Modeling Occlusions and Physical Constraints”. *ICCV*, 2011.
3. Oikonomidis, I., Kyriazis, N., Argyros, A. A. “Efficient Model-based 3D Tracking of Hand Articulations using Kinect”. *BMVC*, 2011.
4. Kyriazis, N., Oikonomidis, I., Argyros., A. A. “A GPU-powered Computational Framework for Efficient 3D Model-based Vision”. *Technical Report TR420, ICS-FORTH*, 2011.
5. Kennedy, J., Eberhart, R. “Particle swarm optimization”. *International Conference on Neural Networks*, 1995.

For more information, visit http://www.ics.forth.gr/~oikonom
or contact {oikonom, kyriazis, argyros}@ics.forth.gr

This work was partially supported by the IST-FP7-IP-215821 project GRASP