Model-based 3D Hand Tracking with on-line Hand Shape Adaptation

Alexandros Makris
amakris@ics.forth.gr
Antonis Argyros
argyros@ics.forth.gr

Institute of Computer Science,
FORTH, Heraklion, Greece
Computer Science Department,
University of Crete,
Heraklion, Greece

3D hand tracking is an interesting problem with high complexity due to the high dimensionality of the human hand and its frequent and often severe self occlusions. Most of the curent methods, track a human hand under the assumption that its parameters (e.g., finger lengths, palm dimensions, e.t.c.) are already known. This assumption limits the applicability of tracking methods. Recently, a few approaches that attempt to solve the hand shape estimation problem have been proposed [5, 6].

In this work we present an on-line method that solves simultaneously the hand tracking and hand shape estimation problems. Let $\mathbf{x}_t$ and $\mathbf{y}_t$ be the pose and shape state at time step t respectively. For the pose estimation the Bayesian Hierarchical Model Framework (HMF) is employed [3, 4]. The framework uses six *auxiliary models* that lie in lower dimensional spaces as proposals for the 26-DOF *main model* of the hand. The shape estimate at each time step is provided by per-frame shape parameters optimization, followed by a robust fitting framework. The per-frame optimizer generates possible shape proposals $\mathbf{y}_t^{pso}$ by optimizing the shape parameters at each frame given fixed (i.e., already estimated) pose parameters. Since the actual shape parameters are constant, the robust fitting cross-validates the shape proposals over a frame history. The output of the fitting is the best estimate given the considered history of the shape parameters $\bar{\mathbf{y}}_t$ that is used in the subsequent frame by the pose tracker.

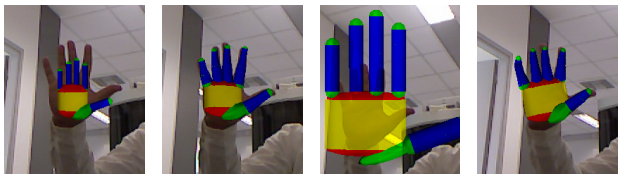
Hand Model The shape of the hand $\mathbf{y}_t$ is parametrized by an 11D vector that controls finger lengths and widths and the width and height of the palm. The pose of the hand $\mathbf{x}_t$ is parametrized by a 27D vector. The kinematics of each finger are modeled using four parameters, two for the base angles and two for the remaining joints. The global position of the hand is represented by a fixed point on the palm and the global orientation.

Pose Tracking The HMF tracking framework [3, 4] that is used to track the hand pose updates at each frame t the pose parameters $\mathbf{x}_t$ given the estimate of the shape parameters $\bar{\mathbf{y}}_{t-1}$. The HMF uses several auxiliary models that are able to provide information for the state of the main model which is to be estimated. Each of the auxiliary models tracks a distinct part of the hand; we use one for the palm with 6-DOF for its 3D position and orientation and one for each finger with 4-DOF for the joint angles. A particle filter is used to sequentially updates the sub-states.

Shape Optimization At each time step t the particle filter described above maintains a set of N weighted particles for the main model. An optimization of the shape parameters using the PSO algorithm is performed independently for the $N^{pso} << N$ particles with the higher weights resulting in N^{pso} updated estimates for the shape parameters paired with the corresponding pose parameters. The likelihood of these pairs is calculated and the shape parameters with the max-likelihood $\mathbf{y}_t^{pso}$ are retained as the current shape estimate.

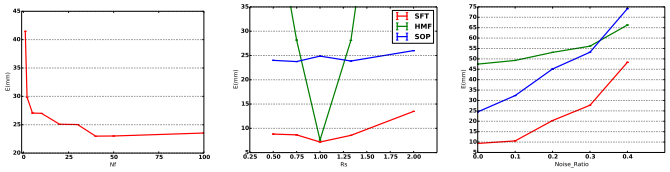
Shape Fitting The per-frame shape estimates up to the current frame are processed by a robust fitting framework. The framework stores a history of N_f frames along with their corresponding poses $H_F = \{\mathbf{z}_f, \bar{\mathbf{x}}_f\}_{f=1}^{N_f}$, and a history of N_s shape parameters $H_S = \{\mathbf{y}_s^{pso}\}_{s=1}^{N_s}$. Every shape $\mathbf{y}_s$ in history H_S is paired with every pose $\bar{\mathbf{x}}_f$ in history H_f. The likelihood $L([\bar{\mathbf{x}}_f, \mathbf{y}_s^{pso}], \mathbf{z}_f)$ of each pair is evaluated and the shape parameters are ranked according to that likelihood. The per-frame ranks $R_f(\mathbf{x}_f, \mathbf{y}_s)$ of each shape parameter set $\mathbf{y}_s^{pso}$ are then averaged to obtain the global rank for the set $R(\mathbf{y}_s)$. The new estimate for the shape parameters is selected by choosing the estimate with the best average rank among the history frames.

Experiments We used real data obtained by RGB-D sensors to qualitatively evaluate the methods and synthetic data for quantitative evaluations. The methods that have been included in our comparative evaluation are: (i) **HMF**: The method of [4] that tracks a hand without estimating its shape. (ii) **SOP**: Tracking the hand through HMF and perform only shape optimization per frame. (iii) **SFT**: The full proposed method. The synthetic dataset that we used for the evaluation consists of 1400 frames of free hand movement. For the shape initialization we test different parameter sets that are scaled with respect to the groundtruth shape by a ratio R_s. We test values for R_s from 0.5 to 2. The pose error E_p measures the average distance between corresponding phalanx endpoints over a sequence similarly to [2]. We are also interested in assessing the performance of the proposed method with respect to noise in the observations. We simulate the imperfect data that come from a real depth sensor, the imperfect foreground detection and the possible occlusions from unknown objects. The results show that the shape estimate converges fast, and this significantly improves the overall tracking accuracy. Test runs on real data are provided in: `https://youtu.be/4dgwoKkDSn8`.



(a) (b) (c) (d)

Figure 1: **SFT** Tracked sequences examples. Two sets of two frames of the same sequence tracked with two different shape initializations. Figures (a) and (c) show the initialization while (b) and (d) the pose/shape estimation several frames later.



(a) (b) (c)

Figure 2: **Quantitative Experiments (a)** Pose error for various maximum history frame values N_f. **(b)** Pose error for various shape initializations. The initialization ratios (x-axis) express the ratio between the shape parameters values that were used for initialization and the ground truth shape parameters. **(c)** Pose error for various sequence noise levels.

Acknowledgments: This work was supported by the EU IST-FP7-IP-288533 project RoboHow.Cog.

[1] I. Albrecht, J. Haber, and H.P. Seidel. Construction and animation of anatomically based human hand models. In SCA, 2003.

[2] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool. Tracking a hand manipulating an object. In *ICCV*, 2009.

[3] A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis. A hierarchical feature fusion framework for adaptive visual tracking. *Image and Vision Computing*, 29(9):594–606, 2011.

[4] A. Makris, N. Kyriazis, and A. Argyros. Hierarchical Particle Filtering for 3d Hand Tracking. In *CVPR, HANDS Workshop*, 2015.

[5] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, and S. Izadi. Accurate, Robust, and Flexible Real-time Hand Tracking. In CHI, NY, USA, 2015. ACM.

[6] J. Taylor, R. Stebbing, V. Ramakrishna, C. Keskin, J. Shotton, S. Izadi, A. Hertzmann, and A. Fitzgibbon. User-specific hand modeling from monocular depth sequences. In *CVPR*, 2014.