Fast and Robust Hand Tracking Using Detection-Guided Optimization

Srinath Sridhar¹, Franziska Mueller¹,², Antti Oulasvirta³, Christian Theobalt¹
¹Max Planck Institute for Informatics. ²Saarland University. ³Aalto University.

There is increasing interest in using markerless hand tracking in human-computer interaction, for instance when interacting with 3D applications, augmented reality, smart watches, and for gestural input [2, 3, 9]. However, flexible, realtime markerless tracking of hands presents several unique challenges. First, natural hand movement involves simultaneous control of several (≥ 25) degrees-of-freedom (DOFs), fast motions with rapid changes in direction, and self-occlusions. Tracking fast and complex finger articulations combined with global motion of the hand at high frame rates is critical but remains a challenging problem. Second, many methods use dense camera setups [4, 8] or GPU acceleration [5], i.e. have high setup costs which limits deployment. Finally, applications of hand tracking demand tracking across many camera-to-scene configurations including desktop, egocentric and wearable settings.

This paper presents a novel method for hand tracking with a single depth camera that aims to address these challenges. Our method is extremely fast (nearly equalling the capture rate of the camera), reliable, and supports varying close-range camera-to-hand arrangements including desktop, and moving egocentric (camera mounted to the head).

The main novelty in our work is a new detection-guided optimization strategy that combines the benefits of two common strands in hand tracking research—model-based generative tracking and discriminative hand pose detection—into a unified framework that yields high efficiency and robust performance and minimizes their mutual failures (see Figure 1). The first contribution in this strategy is a novel, efficient representation of both the input depth and the hand model shape as a mixture of Gaussian functions. While previous work used primitive shapes like cylinders [4, 5] or spheres [6] to represent the hand model, we use Gaussian mixtures for both the depth data and the model. This compact, mathematically smooth representation allows us to formulate pose estimation as a 2.5D generative optimization problem in depth. We define a new depth-only energy, that optimizes for the similarity of the input depth with the hand model. It uses additional prior and data terms to avoid finger collisions and preserve the smoothness of reconstructed motions. Importantly, since the energy is smooth, we can obtain analytic gradients and perform rapid optimization. While pose tracking on this energy alone could run in excess of 120 fps using gradient-based local optimization, this often results in a wrong local pose optimum.

The second contribution in our strategy is thus to incorporate evidence from trained randomized decision forests that label depth pixels into predefined parts of the hand. Unlike previous purely detection-based approaches [1, 7], we use the part labels as additional constraints in an augmented version of the aforementioned depth-only energy, henceforth termed detection-guided energy. The part labels include discriminative detection evidence into generative pose estimation. This enables the tracker to better recover from erroneous local pose optima and prevents temporal jitter common to detection-only approaches. The precondition for recovery is reliability of the part labels. However, even with large training sets it is hard to obtain perfect part classification (per-pixel accuracy is usually around 60%). Thus, pose estimation based on this additional discriminative evidence is also not sufficient.

Our third contribution therefore, is a new late fusion approach that combines particle-based multi-hypothesis optimization with an efficient local gradient-based optimizer. Previous work has used particle-based optimizers, but they tend to be computationally expensive [4, 5]. Our approach is fast because we combine the speed of local gradient-based optimization with the robustness of particle-based approaches. At each time step of depth video, a set of initial pose hypotheses (particles) is generated, from which a subsequent local optimization is started. Some of these local optimizers use the depth-only pose energy, some others use the detection-guided energy.

In a final late fusion step the best pose is chosen based on the pose fitting energy.

Our approach results in a temporally stable and efficient tracker that estimates full articulated joint angles of even rapid and complex hand motions at previously unseen frame rates in excess of 50 fps, even with a CPU implementation. Our tracker is resilient to erroneous local convergence by resorting to the detection-guided solution when labels can be trusted, and it is not misled by erroneous detections as it can then switch to the depth-only tracking result.

We show these improvements with (1) qualitative experiments, (2) extensive evaluation on public datasets, and (3) comparisons with other state-of-the-art methods.

Figure 1: We develop a novel representation for depth data and hand model as a mixture of 2.5D Gaussians. This representation allows us to combine the benefits of model-based generative tracking and discriminative part detection. Pixels classified using a trained decision forest are directly incorporated as evidence in detection-guided pose optimization. Dashed lines indicate offline computation. Best viewed in color.