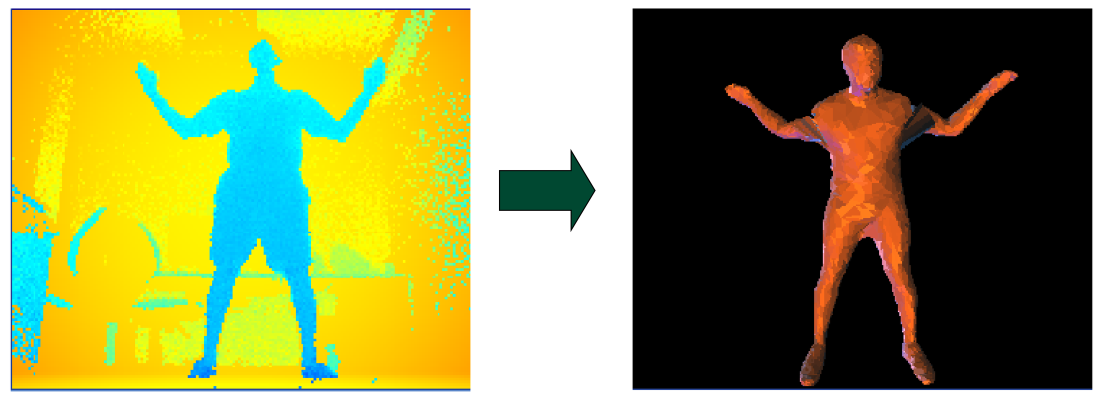
Stanford University
3D Vision Lab
3d.stanford.edu

GPU Accelerated Marker-less Motion Capture

Varun Ganapathi Christian Plagemann Daphne Koller Sebastian Thrun

Goal

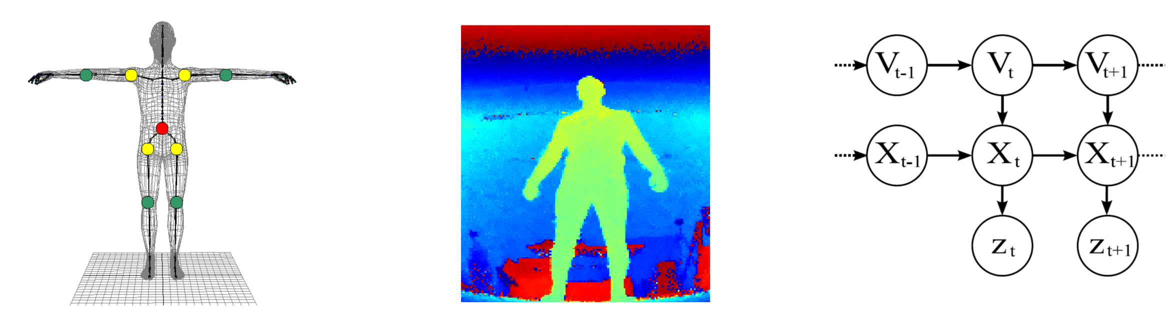


Depth images at 30hz Full 3D pose in real time

Given a sequence of depth images of a human subject, estimate the 3D locations of all joints in real time (shoulder, knee, etc.)

Potential Applications:
Human-machine interaction, smart surveillance, ani-mation, virtual reality and motion analysis.

Probabilistic Model



Kinematic chain z_t : Depth image DBN

X_t Relative poses of all N body parts at time *t*
V_t First time derivative of pose
X^i Pose of part i relative to its parent
z_t Depth image at time t

State transition model assumes random accelerations and that the state is a deterministic function of velocity and the previous state:

$$V_t | V_{t-1} \sim \mathcal{N}(V_{t-1}, \Sigma) \qquad X_t^i = V_t^i X_{t-1}^i$$

Sensor model assumes that the range scan is generated by ray casting. Each pixel k is therefore conditionally independent given the pose and mesh:

$$P(z_t | X_t, m) = \prod_k P(z_t^k | X_t, m)$$

The depth at pixel k is generated by rendering a skinned mesh to calculate the true distance z*

Smooth likelihood by allowing the ray to hit a neighboring pixel in rendered depth scan.

Likelihood can be evaluated efficiently on a GPU by using shaders for differencing the measured data and rendered pose. Use glGenerate-MipMaps for computing the average pixel error.

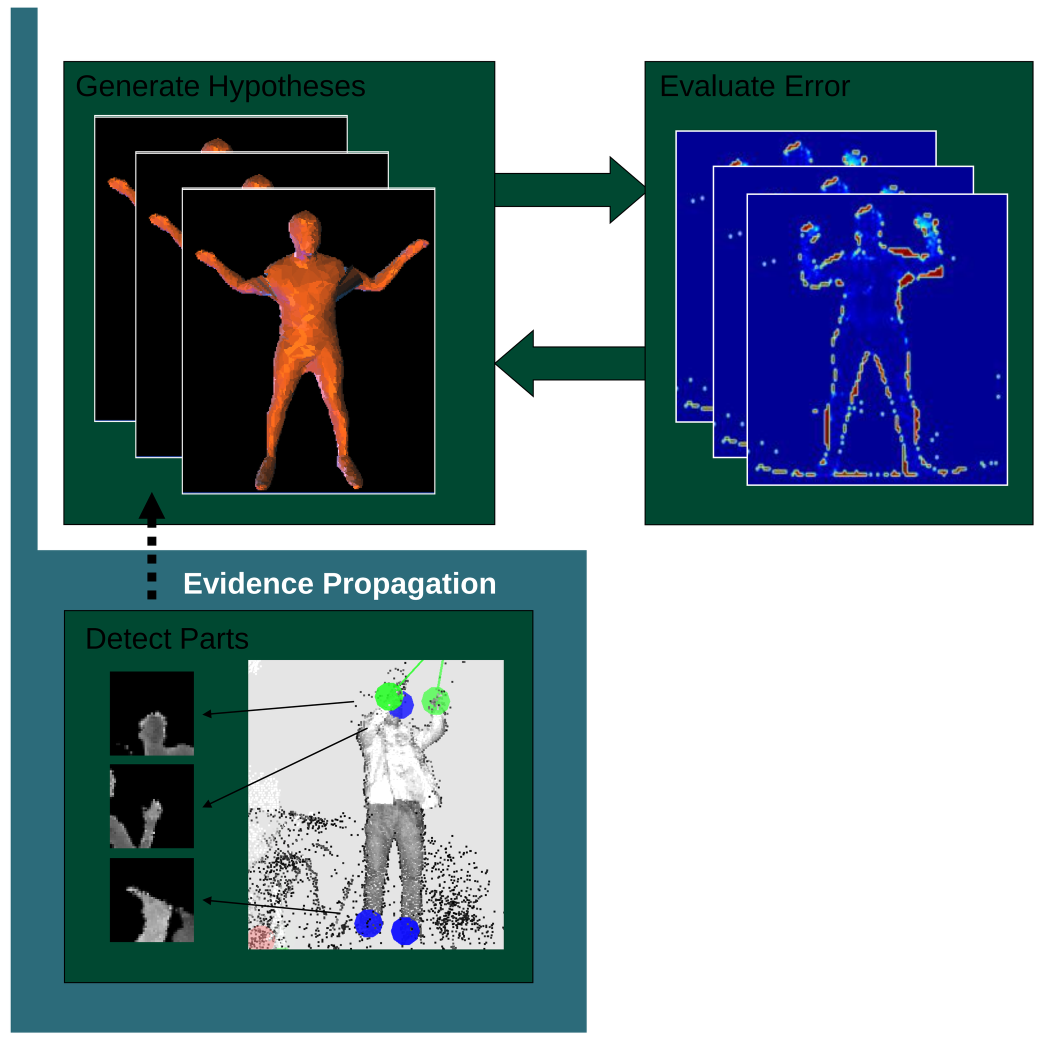
Inference

Objective: Find the body pose that maximizes the posterior likelihood of the observed depth images:

$$\operatorname{argmax}_{X_t, V_t} \log P(z_t | X_t, V_t) + \log P(X_t, V_t | \hat{X}_{t-1}, \hat{V}_{t-1})$$

Challenge: High-dimensional state (48 DoF) and non-linear, noisy dependencies; Real-time constraints.

Our Approach: GPU-accelerated local hill climbing in model space + integration of body part detections to track fast and difficult motions.



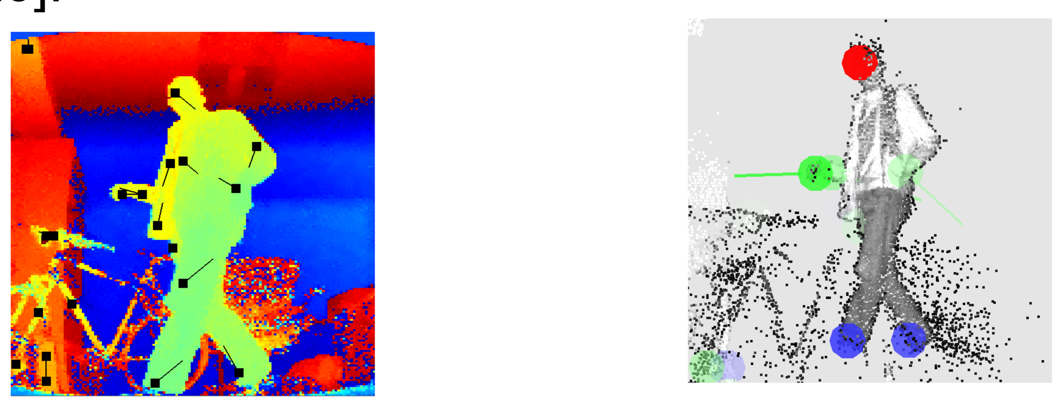
Local Hill Climbing

Starting from the root of the kinematic chain, we sample perturbations to the state. For each dimension, we sample using a coarse grid of joint angle perturbations, followed by a finer sampling.

Large batches of pose hypotheses are evaluated simultaneously using the GPU.

Body Part Detection

Extract interest points from the surface mesh and assign body part labels [Plagemann *et al.*, ICRA 2010].

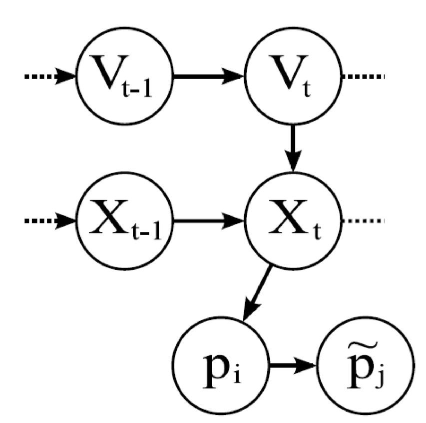


$\tilde{c}_j$ Class label, one of { head, hand, foot }
$\tilde{\mathbf{p}}_j$ 3d location
$\{\tilde{\mathbf{p}}_j, \tilde{c}_j\}$ List of part detection candidates

Evidence Propagation

Integrate body part detections into the set of pose hypotheses.

Auxiliary probabilistic model relating associated detections to the state:



$\mathbf{p}_i$ position of model vertex i as a function of pose
$\tilde{\mathbf{p}}_j$ corresponding body part
$\tilde{\mathbf{p}}_j \sim \mathcal{N}(\mathbf{p}_i(X), \Sigma_o)$

Problem: pi ~ X is heavily non-linear. The location of a vertex p_i is a function of the part it is located in and the pose of the part. W^i the pose of part i a product of the poses of its ancestors:

$$W^i(X) = X^1 \cdots X^{\mathrm{parent}(i)} X^i$$

Our Solution: Apply the unscented transform to linearize about the current state. This results in a linear Gaussian network, in which MAP inference is easy. The procedure can be repeated until convergence.

Data Association

Problem: Detections consist only of location and class. How to associate detections to model vertices, and reject false detections? Exponential # of possibilities!

Our Solution: Prune associations explained by current estimate. Consider associations one at time and accept those that improve the likelihood when integrated using EP.

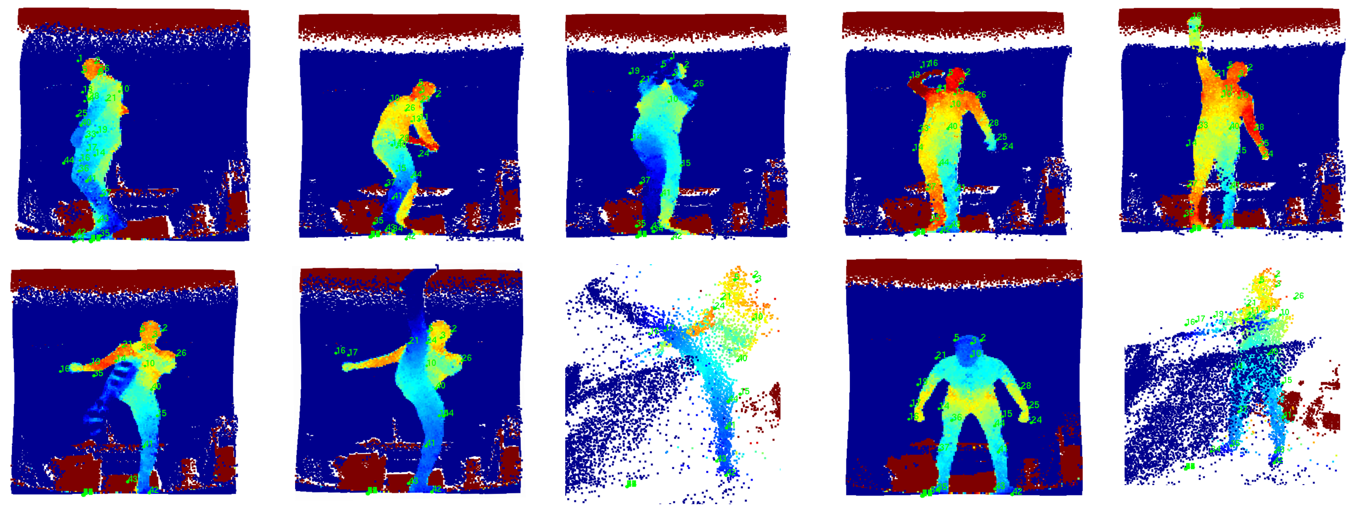
Complete Algorithm

1. Update X^{best} by local hill-climbing on the likelihood
2. Extract part detections from z_t
3. Prune hypotheses that are already explained
4. Produce $\boxed{N}$ correspondences $\{(\mathbf{p}_i, \tilde{\mathbf{p}}_j)\}$ by expand-ing hypotheses
5. Loop $i = 1$ to N
   (a) Let X' be the posterior mode of evidence propa-gation initialized from X^{best} conditioned on c^i
   (b) Update X' by local hill-climbing on likelihood
   (c) if likelihood of $X' > X^{\text{best}}$, set X^{best} to X_c

Experiments

Dataset

28 sequences of various difficulty. We simultaneously recorded marker location traces use an active marker system along with frames of depth data at 25 FPS using the SR4k. Sample frames:



Results

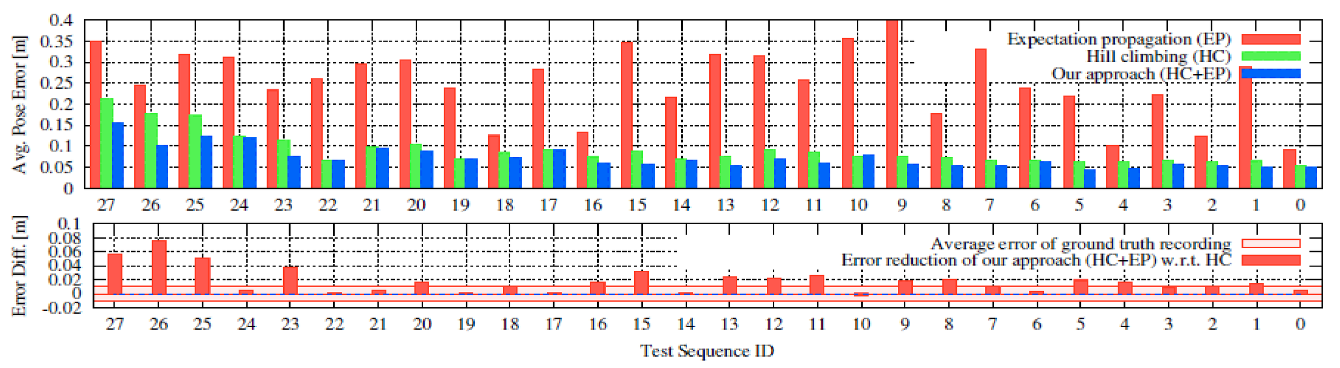


Figure 5. Tracking results on real-world test sequences, sorted from most complex (left) to least complex (right).

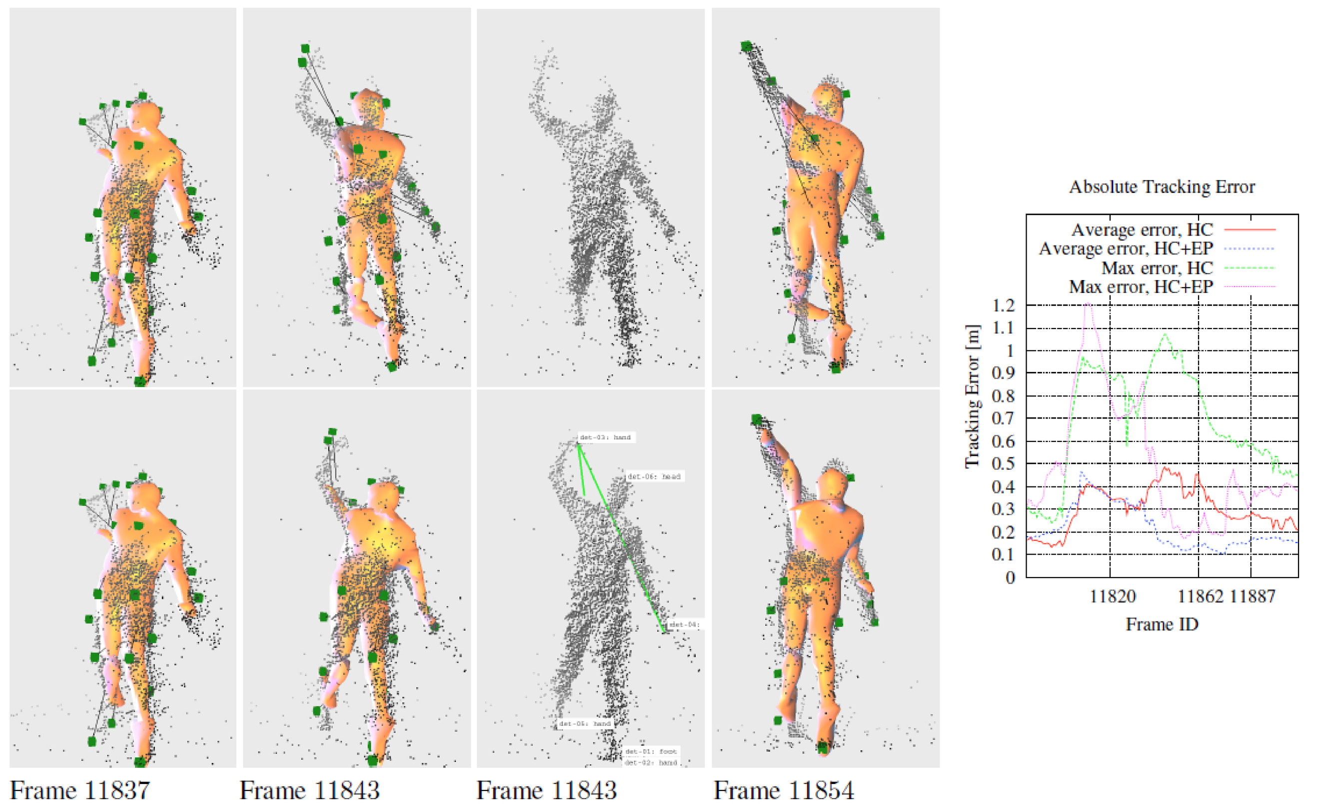


Figure 6. A typical situation in which data-driven evidence is crucial for tracking success (excerpt from Seq. 27). Left: Three exemplary frames from the *Tennis* sequence. *Model-Based search* (top row) loses track of the tennis swing, since the arm was occluded. Our combined tracker that integrates bottom-up evidence about body parts (bottom row) is able to recapture the fast moving arm. The right diagram shows the same situation in terms of actual tracking error (see text).

Conclusion

With the hybrid generative /discriminative GPU-accelerated filtering approach introduced in this paper, we believe to have made a large step forward, but there remain more challenges to overcome.

Some examples include cluttered scenes, multiple people, automatic model initialization, improved speed and robustness. Extremely fast motions remain difficult to track with current sensors.