Capturing Hand Motion with an RGB-D Sensor, Fusing a Generative Model with Salient Points

Supplementary Material

Dimitrios Tzionas[1,2], Abhilash Srikantha[1,2], Pablo Aponte[2], and Juergen Gall[2]

[1] Perceiving Systems Department, MPI for Intelligent Systems, Germany
dimitris.tzionas@tue.mpg.de, abhilash.srikantha@tue.mpg.de
[2] Computer Vision Group, University of Bonn, Germany
aponte@iai.uni-bonn.de, gall@iai.uni-bonn.de

1 Hand Model

For the pose estimation we resort to the widely used Linear Blend Skinning model [4], consisting of a triangular mesh, an underlying kinematic skeleton and a set of skinning weights. A personalized model for a single subject was created with the following rigging process:

A detailed triangular mesh of both hands of the subject was created using a commercial 3D scanning solution[1]. The scanning setup consisted of 5 camera-pods[2] placed in proximity to the scanned hand, covering several viewpoints. A 3D mesh was reconstructed with the camera-system's proprietary software (multi-view stereo), which was further denoised and processed (e.g. hole filling) manually using Meshlab[3]. The final result was a watertight mesh for each hand consisting of approximately 10.000 vertices. A skeletal structure was manually fitted in each mesh using a custom OpenGL tool and the corresponding skinning weights for each vertex were computed using the open-source[4] "Pinocchio" software [2]. In our experiments, a single hand consists of 31 revolute joints, i.e. 37 DoF (including 6 DoF for the global rigid motion that models the wrist). Thus, for sequences with two interacting hands we have to estimate all 74 DoF. Figure 1 depicts the mesh, the skeleton and the DoF for the right hand.

2 Sequences

We evaluate our approach, both qualitatively and quantitatively, on 14 new manually annotated[5] sequences with challenging interactions. Table 1 notes detailed information about these sequences and the way they were used in our experiments. *Set A*, including 11 sequences, is used in order to evaluate the components

[1] http://www.3dmd.com
[2] Each pod consists of 1 RGB camera, 2 B&W cameras and 2 speckle projectors.
[3] http://meshlab.sourceforge.net
[4] http://www.mit.edu/~ibaran/autorig/pinocchio.html
[5] Annotation takes place every 5^{th} frame.

of our pipeline, while *Set B*, including the remaining 3, is used for comparison with the state-of-the-art method [5]. The *starting* and *ending* frame ID that was used for the tracker is noted, while each sequence is characterized by the intensity of observed collisions.

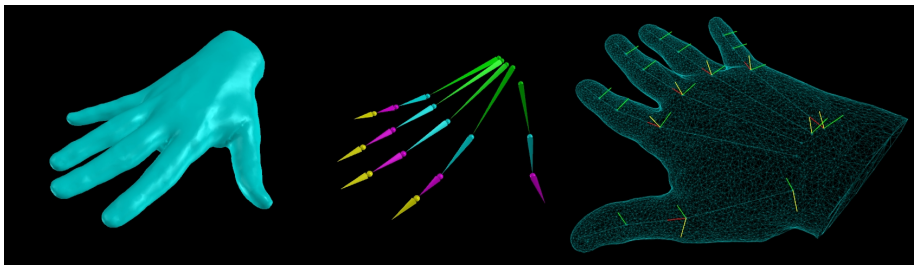
3 Salient Point Detector

Our method is primarily based on local optimization and a generative model, which generally provides accurate solutions. However, its accuracy depends on good initialization and is prone to accumulative error and local minima. A discriminative method can be an effective complement in that matter, driving the optimization framework away from local minima in the search space and aiding convergence to the global minimum.

For this reason we employ discriminatively trained salient points (fingertips) on raw depth images using a Hough forest [3]. We annotated a set of 56 sequences consisting of approximately 2000 frames. The Hough forest consists of 17 trees and has maximum tree-depth of 25 nodes. The features are computed solely on raw depth data, based on patches of 16×16 size. The video **fingertipAnnotations.mp4** depicts the annotated training- and test-set. These sets were created having a generic fingertip detector in mind, including hands in isolation, interacting hands and hand-object interaction.

During tracking, correspondences need to be found between detections and mesh-fingertips. We refer to fingertip vertices as *Vertices of Interest* (VOI), which are depicted in Figure 2 (green color). We manually define for each finger one *source-VOI* (red vertex in Figure 2), based on which a set of VOIs can be found. The centroid of the VOIs (yellow vertex in Figure 2) is used in order to compute the 3D distance between each fingertip and each detection (using their 3D point cloud centroid).

The video **detections.mp4** depicts color-coded detection to fingertip associations for the sequence "Helix - Blend".



(a) (b) (c)

Fig. 1. Model used for tracking. (a) Mesh (b) Skeleton (c) Degrees of Freedom (DoF)

Table 1. Sequences. *Set A* is used for evaluation of the components of the presented pipeline, while *Set B* is used as a comparison benchmark with the FORTH tracker [5]. All frames of *Set A* are used for evaluation, while for the sequences of *Set B* the evaluation starts at the noted *starting frame* ("ID Start"), since initialization of the compared trackers is different, while the last frame is rejected, since the public software of [5] failed for the last frame of one sequence. The number of the hands in each scene is noted, as well as the characterization of the collisions that take place in the scene: *some*, *severe* and *no apparent* collision. Only two hand sequences can be characterized by *severe* collisions. The public software of [5] can handle tracking of only one hand

	Sequence	*ID*	*Hands*	*Total*	*ID Start*	*ID End*	*Collision*
Set A	*Walk*	1	2	231	0	$total-1$	*Severe*
	Cross	2	2	153	0	$total-1$	*Severe*
	Cross & Twist	3	2	155	0	$total-1$	*Severe*
	Helix - Tips	4	2	173	0	$total-1$	*Some*
	Dance	5	2	265	0	$total-1$	*Severe*
	Helix - Blend	6	2	136	0	$total-1$	*No*
	Hug	7	2	194	0	$total-1$	*Severe*
	Grasp	8	1	106	0	$total-1$	*No*
	Fly	9	1	138	0	$total-1$	*No*
	Rock	10	1	139	0	$total-1$	*Some*
	Bunny	11	1	134	0	$total-1$	*Some*
Set B	*Bunny*	12	1	727	420	$total-2$	*Some*
	Fly	13	1	778	480	$total-2$	*No*
	Rock	14	1	378	250	$total-2$	*Some*

4 Qualitative Results - Our tracker

Qualitative results of our pipeline for the sequences of *Set A* presented in Table 1 are shown in two videos. Results and input images for all sequences are shown in **ourResults.mp4**, while results and input point clouds are shown in **ourResultsPCL.mp4**.

5 Quantitative Results - Ground Truth Annotation

The sequences were manually annotated by four subjects in order to enable quantitative analysis and comparisons. The annotation takes place every fifth frame, while each sequence is annotated by a single annotator.

In order to test the annotation aquracy, 10% of the total frames is sampled and annotated by all four subjects. The standard deviation for all subjects over all frames and joints is 1.46 pixels.

6 Quantitative Results - Error Metric

The error metric used for the evaluation of our pipeline and the comparison against the state-of-the-art is the 2D distance (pixel units) between the projec-

Fig. 2. Vertices of Interest (VOI) (green) used for the salient point detector. The red vertex is the *source* VOI, which is manually defined. The centroid of the VOIs, which is important for the assignment of detections to fingers, is depicted with yellow

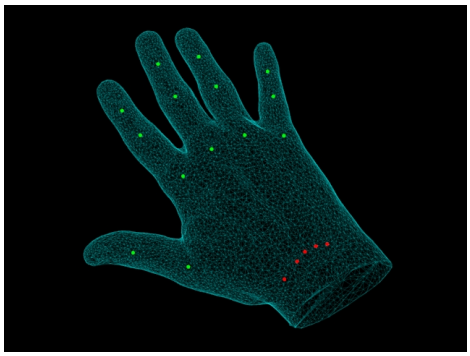


Fig. 3. Hand joints used in the error metric. The 14 joints taken into consideration are depicted with green color, while the ignored ones are depicted with red color

tion of the 3D joints and the corresponding 2D annotations. Figure 3 shows the hand joints taken into account for the computation of the error metric. Joints around the wrist are ingored because their annotation can be very noisy.

7 Quantitative Results - Comparison to State-of-the-Art

Recently, Oikonomidis et al. used Particle Swarm Optimization (PSO) for a real-time hand tracker [5, 6, 7]. These works constitute the state-of-the-art for single-view RGB-D hand tracking. For comparison, we use the software[6] released for tracking one hand [5], with the parameter setups of all the above works (for the later works [6, 7] no software was released for comparisons). Each setup is evaluated 3 times in order to compensate for the manual initialization and the inherent randomness of PSO. This process is visualized in the

[6] `http://cvrlcode.ics.forth.gr/handtracking`

Table 2. Comparison of our method against the FORTH tracker. The FORTH tracker is evaluated with 4 parameter setups met in the referenced literature of the last column

		Repeats per Set			Average			Reference
		Mean (px)	*St.Dev. (px)*	*Max (px)*	*Mean (px)*	*St.Dev (px)*	*Max (px)*	
FORTH	*set 1*	8.44646	5.82398	57.6281	8.58347	5.74316	61.8142	[5]
		8.51244	5.33565	37.3363				
		8.79152	6.04124	61.8142				
	set 2	8.05772	5.04593	36.0694	8.32383	5.42152	57.9741	[6]
		8.94478	6.15262	57.9741				
		7.96900	4.92750	36.4005				
	set 3	8.13842	4.82596	33.1059	8.08535	5.00020	38.8973	[1]
		8.15380	5.17905	32.3883				
		7.96383	4.98710	38.8973				
	set 4	8.15469	5.24789	39.8497	8.15868	5.17618	39.8497	[7]
		8.21784	5.17665	37.4433				
		8.10351	5.10235	36.6197				
Our		3.75551	2.21604	19.9249				

video named **benchmarkFORTH.mp4**. The initialization process is shown for the parameter set 1 of the "Bunny" sequence (ID 12) in the file **benchmarkFORTH_FullBunnySet1.mp4**. Quantitative results of Table 2 show that our system outperforms [5] in terms of tracking accuracy. A qualitative comparison of all our results with the best version of each parameter setup of [5] is included in the video named **comparisonFORTH.mp4**.

8 Collision Detection

A video (**collisionDetection.mp4**) showcasing the collision detection component in action is included for a part of the sequence "*Walk*", which includes *severe* collisions (Table 1). All stages of the tracker are shown and the colliding triangles are depicted with red color.

9 Runtime

The runtime of our pipeline was measured with one sequence per scene complexity: sequence "Cross and Twist" (ID 3) which contains 2 hands and sequence "Bunny" (ID 11) which contains 1 hand. It refers to unoptimized single-threaded code running on an Intel Core i7-4930K CPU. GPU is used just for rendering. The runtime for the chosen setup, as described in the paper, is 2.74 and 4.35 seconds per frame for scenes containing one and two hands respectively.

References

1. Ballan, L., Taneja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: ECCV. pp. 640–653 (2012)
2. Baran, I., Popović, J.: Automatic rigging and animation of 3d characters. TOG 26(3) (2007)
3. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.: Hough forests for object detection, tracking, and action recognition. PAMI 33(11), 2188–2202 (2011)
4. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In: SIGGRAPH. pp. 165–172 (2000)
5. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3d tracking of hand articulations using kinect. In: BMVC. pp. 101.1–101.11 (2011)
6. Oikonomidis, I., Kyriazis, N., Argyros, A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: ICCV. pp. 2088–2095 (2011)
7. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the articulated motion of two strongly interacting hands. In: CVPR. pp. 1862–1869 (2012)