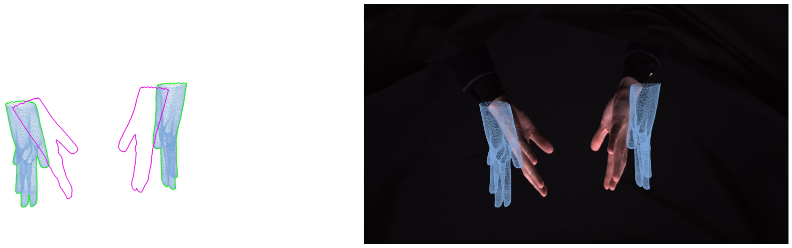
A Comparison of Directional Distances for Hand Pose Estimation **Supplementary Material***

Dimitrios Tzionas[1,2] and Juergen Gall[2]

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

Pictorial Description of the Benchmarking Test Pairs

The proposed benchmarking protocol analyzes the error not over full sequences, but over a sampled *set of testing pairs*, consisting of a *starting pose* and a *test frame*. In this respect, 4 publicly available sequences[1] are used, containing realistic scenarios of two strongly interacting hands. 10% of the total frames are randomly selected, forming the set of *test frames* of the final pairs. This is the basis to create 4 different sets of image pairs, having 1,5,10,15 frames difference, respectively, between the *starting pose* and the *test frame*, presenting thus increasing difficulty for tracking systems. These 4 sets and the overall combination constitute a challenging dataset, representing realistic scenarios the occur due to low frame rates, fast motion or estimation errors in the previous frame. The created testing sets are used in two experimental setups: a purely *synthetic* and a *realistic*.

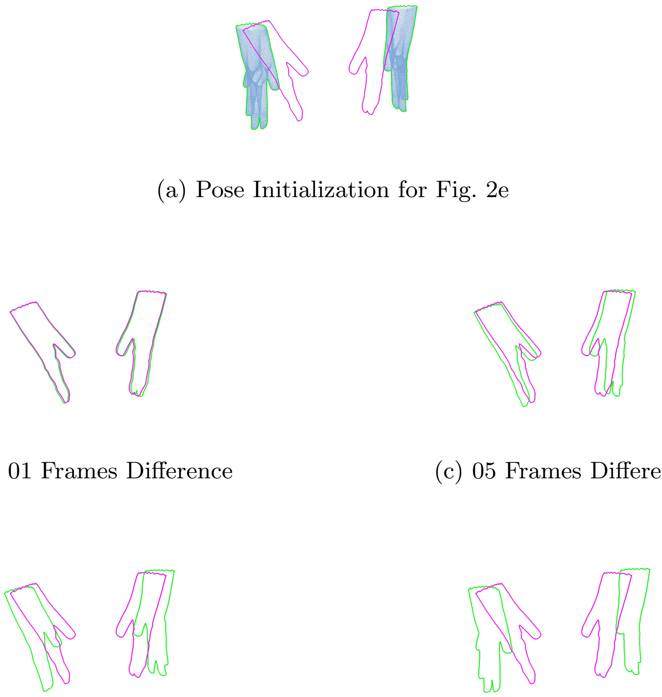


(a) Synthetic

(b) Real

Fig. 1: Examples of *pose initialization* for the exhibited set of test frames, for both the synthetic and real experimental setup. The silhouette of the starting pose in following figures is created by the projection of the mesh after this initialization. In Fig. 1a the *silhouette of the starting pose* is depicted with green color, while the *target silhouette* is depicted with magenta.

[1] Model, videos, and motion data are provided at http://cvg.ethz.ch/research/ih-mocap. Sequences: *Finger tips touching and praying, Fingers crossing and twisting, Fingers folding, Fingers walking.* Video: 1080 × 1920 px, 50 fps, 8 camera-views.



(a) Pose Initialization for Fig. 2e

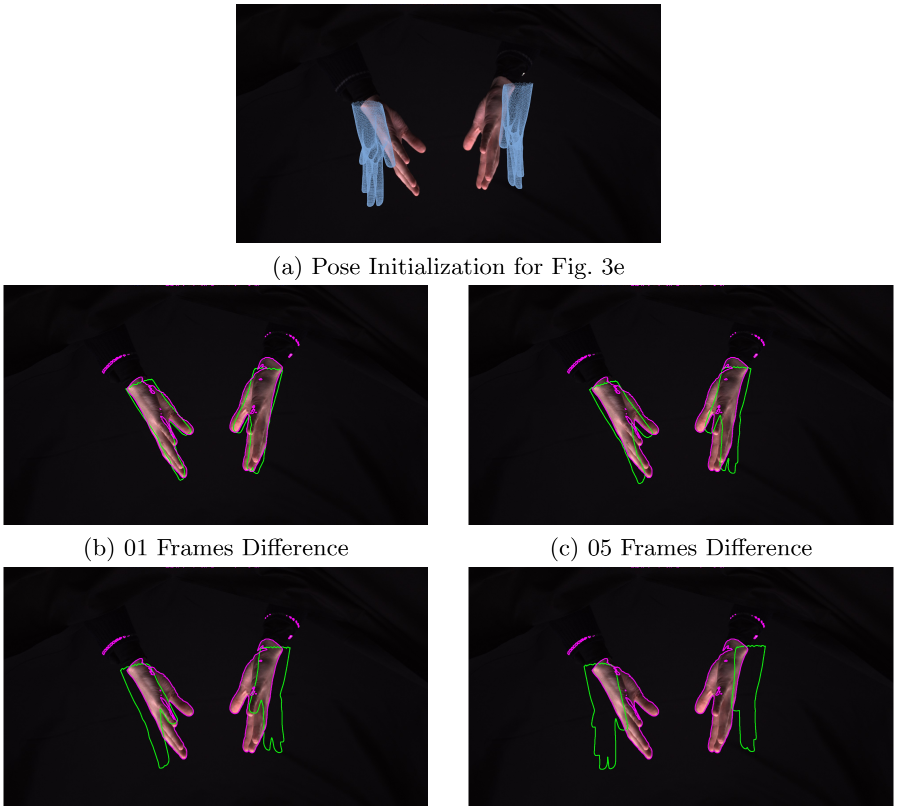
(b) 01 Frames Difference

(c) 05 Frames Difference

(d) 10 Frames Difference

(e) 15 Frames Difference

Fig. 2: Test pairs in the case of **synthetic** experiments. The *test frame* is defined by sampling the whole set of sequence frames, while the *starting pose* is chosen in such a way, so that 4 different pairs are created, with frame difference of 1,5,10,15 frames. The *silhouette of the starting pose* is depicted with green color, while the *target silhouette* is depicted with magenta.



(a) Pose Initialization for Fig. 3e

(b) 01 Frames Difference

(c) 05 Frames Difference

(d) 10 Frames Difference

(e) 15 Frames Difference

Fig. 3: Test pairs in the case of **real** experiments. The *test frame* is defined by sampling the whole set of sequence frames, while the *starting pose* is chosen in such a way, so that 4 different pairs are created, with frame difference of 1,5,10,15 frames. The *silhouette of the starting pose* is depicted with green color, while the *target silhouette* is depicted with magenta.

Manual Annotation of 3d Hand Joints

For the manual annotation of the 3d ground-truth hand joints by a human subject, the following process was followed:

The subject had to mark in a predefined order all the visible joints of the two hands for all camera views and the 2d coordinates of the joints where stored. In case of occlusion or high ambiguity, the 2d joint position for the respective camera view was marked as missing/ambiguous and it was not taken into account. In order to reconstruct the 3d ground-truth position of a joint, a 3d projection ray was computed for each 2d joint projection of it. For all possible couples of projection rays the 3d point that minimizes the distance to both rays was computed. The final ground-truth 3d coordinates of each joint where given by averaging the aforementioned corresponding 3d points.