Low-Dimensionality Calibration through Local Anisotropic Scaling for Robust Hand Model Personalization

Edoardo Remelli* EPFL
Anastasia Tkach* EPFL
Andrea Tagliasacchi University of Victoria
Mark Pauly EPFL

Abstract

We present a robust algorithm for personalizing a sphere-mesh tracking model to a user from a collection of depth measurements. Our core contribution is to demonstrate how simple geometric reasoning can be exploited to build a shape-space, and how its performance is comparable to shape-spaces constructed from datasets of carefully calibrated models. We achieve this goal by first re-parameterizing the geometry of the tracking template, and introducing a multi-stage calibration optimization. Our novel parameterization decouples the degrees of freedom for pose and shape, resulting in improved convergence properties. Our analytically differentiable multi-stage calibration pipeline optimizes for the model in the natural low-dimensional space of local anisotropic scalings, leading to an effective solution that can be easily embedded in other tracking/calibration algorithms. Compared to existing sphere-mesh calibration algorithms, quantitative experiments assess our algorithm possesses a larger convergence basin, and our personalized models allows to perform motion tracking with superior accuracy.

1. Introduction

Recent developments in visualization and motion tracking technologies are rapidly bringing virtual and augmented reality products to the mainstream consumer market. Such technologies have a much wider target than the gaming and entertainment industry alone, with new exciting applications in a variety of fields, ranging from education to medical industry. One of the main technical challenges that VR/AR has to face is to be able to offer the user the capability of interacting naturally with the surrounding virtual environment. In real life we interact with objects using our hands: *can hand tracking enable natural and accurate interactions with virtual objects?* This task is particularly challenging, as hands are highly articulated, undergo frequent occlusions and self-occlusions, fingers have similar appearance, and global pose is unconstrained. With the advent of consumer depth cameras, there has been substantial progress towards the development of robust hand trackers [13, 21, 27, 28]. Most approaches leverage a combination of discriminative and generative algorithms, where the former require no temporal history but can provide a rough initialization, the latter can exploit such initialization, temporal coherency, and shape priors to accurately resolve alignment. The better the generative model can fit to the user, the better tracking accuracy can be achieved.

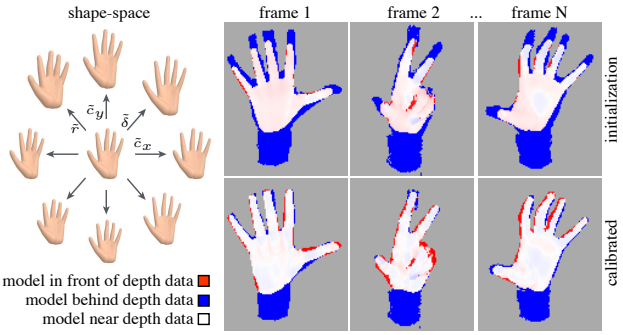

Figure 1: Our low-dimensional latent space directly encodes shape variations as local anisotropic scalings of a template model. (left) A visualization of the shape-space, encoding palm/finger length compression/stretching and sphere-mesh radii. (right) Calibrating in the reduced shape-space followed by a fine scale refinement leads to a robust calibration algorithm – remaining errors are largely due to sensor artefacts.

Robust hand model calibration. State of the art techniques such as [27, 28] achieve high-precision tracking by personalizing the template to the given subject. Given (1) a *set of depth images* and (2) a *sufficiently close initialization*, calibration optimization algorithms can be employed to generate a personalized triangular-mesh [26] or sphere-mesh [28] model of the observed user's hand. Recently, Tan et al. [24] relaxed the necessity to have a tight initialization

*equal contributors

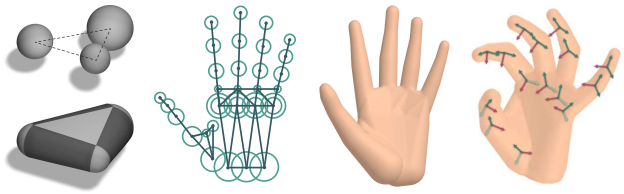

Figure 2: (a) A basic sphere-mesh component whose surface is the convex hull of spheres whose radii are defined on the skeleton vertices. (b) The sphere-mesh hand topology and (c) the rest-pose tracking template. (d) The posed template, and a visualization of the kinematic frames $\{\bar{\boldsymbol{T}}_k\}$.

by proposing a robust calibration that leverages the shape-space introduced by [7] – a *robust* calibration optimization is one exhibiting few local minima, or, analogously, one possessing a wide basin of convergence. The shape-space of [7] regularizes the calibration optimization for use in uncontrolled setups, preventing it from converging towards local minima. It was constructed from depth images collected from a large set of users in a controlled laboratory setup, and calibrating a consistent template to the data [26]. In our research we challenge these assumptions and pose the question: "*Is the construction of a data-driven shape-space strictly necessary to achieve robust optimization?*" We approach this challenge by formulating our calibration optimization as the one of personalizing the *sphere-mesh* tracking template introduced by Tkach et al. [28] to a given user; see Figure 2. While data-driven shape-spaces will still be helpful for calibrating high-frequency details [3], we show how a geometric shape-space is sufficient in calibrating sphere-mesh models to an accuracy that is appropriate to the signal-to-noise ratio of current generation depth sensors.

Contributions. We improve upon the calibration optimization proposed in [28] by addressing its robustness shortcomings, thanks to a novel parameterization for sphere-mesh calibration optimization. In contrast to [28], our parameterization *decouples* pose and shape parameters, resulting in faster convergence and increased robustness; see Section 4. For the calibration algorithm to be sufficiently robust to be effectively embedded in a consumer-level tracking system, inspired by Tan et al. [24], we propose to calibrate in a low dimensional shape-space. However, we do not derive this shape-space from a dataset, which can be cumbersome to acquire, process, and interpret. Rather, we leverage simple geometric observations, and *construct a local anisotropic scaling shape-space*; see Section 3. As opposed to the optimization scheme of Tan et al. [24] relying on numerical differentiation, our calibration algorithm is *analytically differentiable*. We also demonstrate how our shape-space can be easily integrated in the calibration framework of [28], and

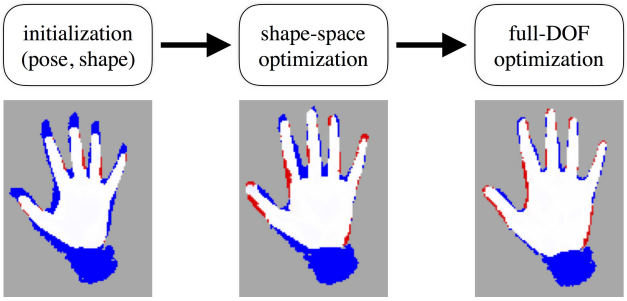

Figure 3: In initialization our calibration algorithm receives the default template tracked by [28] that provides an initialization of pose and shape. We then optimize for a coarse shape in a low-dimensional shape-space, and finally perform a local refinement in the full-dimensional domain.

discuss how it could also easily be integrated in other calibration algorithms based on Levenberg optimization. This is in clear contrast to Tan et al. [24], as the shape-space introduced by Khamis et al. [7] is *specific* to the chosen tracking model, thus requiring re-parameterization prior to adoption in other systems. We integrate all our contributions in a multi-stage calibration algorithm; see Figure 3.

2. Related Work

A large body of work exists for human hand and body tracking. We focus on the former, and refer to the recent work of Bogo et al. [3] for a more complete overview on full-body tracking. As model personalization is a core ingredient in generative motion tracking [14], we cover this topic with more attention. Several approaches for hand tracking have been proposed, and the type of input data has evolved alongside sensing technology. Classical examples of instrumented acquisition include sensored [5] or colored [32] gloves, accelerometers [34, 31], marker-based capture [33, 36] and wearable cameras [8]. While effective, active instrumentation is encumbering and requires complex and time-consuming calibration, a requirement making this approach impractical for consumer-level applications. Similarly, requiring the user to wear a camera, a glove or placing markers disrupts the potential for seamless human-machine interaction. Un-instrumented multi-camera systems such as [2, 19] can lead to excellent tracking quality, but multi-sensor setup and calibration is impractical; further, due to the large amount of data to be processed, real-time performance is difficult to achieve. The use of a *single* sensor, providing either color or depth streams, represents the most logical choice for consumer-level applications. Hence, in our approach we focus on single depth camera input. Modern hand tracking techniques from depth data employ algorithms that are gener-

ative [16, 13, 10, 20, 23, 28], discriminative [30, 11, 12] or a combination of the two [15, 18, 21, 27]. Due to the large body of recent works in this area, we refer the reader to the survey by [22] and recent works [27, 35] for more details.

Calibration and shape-spaces. Albrecht et al. [1] pioneered the construction of realistic (skin, bone and muscles) personalized models through registration of 3D mesh scanned from a plaster. Skin creases and silhouette images can also be used to guide the registration of a model to color imagery [17]. More closely related to ours is the work by de La Gorce et al. [4], where a triangular-mesh template is calibrated to a rest pose hand image. This is achieved by augmenting the nodes of the kinematic tree with (relative) scaling transformations, resulting in a 53 DOF optimization problem. Such a scheme is only capable to adjust finger lengths/girths, while palm geometry cannot be adapted to the user beyond simple scaling. Further, linear constraints bounding (relative) scale magnitudes are necessary to avoid the solution from drifting; these hard constraints limit the expressivity of the model. While few details are available for this algorithm, we speculate such a scheme suffers convergence shortcomings analogous to the full-DOF optimization of Section 3. A different approach was taken by Makris and Argyros [9], who solve for the cylindrical geometry of a hand through render-and-compare evaluations optimized by particle swarm optimization. In contrast to [4, 9, 24], our method relies on *analytical* differentiation, leading the way to the implementation of high-performance calibration routines. Our sphere-mesh model is more expressive than the cylinders in [9] or the scaled model in [4], while being well suited to the signal-to-noise ratio of currently available real-time depth cameras.

3. Robust model calibration

Similarly to recently proposed calibration optimization of [28, 24], we design a routine fitting our sphere-mesh template model $\bar{\mathcal{S}}$ to a collection of N single-view depth measurements $\{\mathcal{D}_1, ..., \mathcal{D}_N\}$ of a specific user's hand in different poses. Redundancy in the measurement is necessary due to the incompleteness of single-view measurements (i.e. at most 50% of the model is visible in any frame), and to the inability to determine certain degrees of freedom in specific configurations (e.g. the length of a phalanx can only be measured when a finger is bent).

Full-DOF optimization. We cast the calibration problem into an optimization over pose and shape by defining the following multi-objective energy function:

$$\underset{\{\boldsymbol{\Psi}_n\}}{\arg\min} \sum_{n=1}^{N} \sum_{\tau \in \tau_{\text{calib}}} \omega_\tau E_\tau(\boldsymbol{\Psi}_n; \mathcal{D}_n) \tag{1}$$

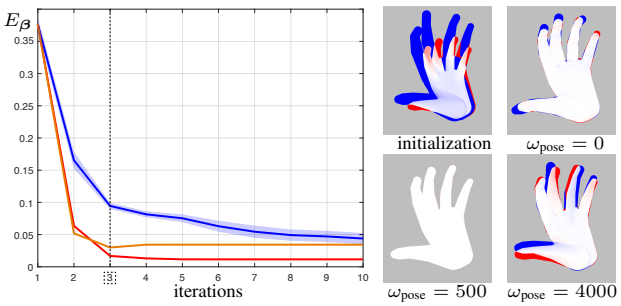

Figure 4: The decoupled parameterization allows to damp updates in pose and shape with different weights, leading to faster and more robust convergence. (left) In the plot above we analyze the convergence for $\omega_{\text{pose}} = \{0, 500, 4000\}$. (right) We illustrate the calibrated model at iteration 3.

where $\boldsymbol{\Psi}_n = \{\boldsymbol{\theta}_n, \boldsymbol{\beta}\}$ encodes the parameters defining the shape $\boldsymbol{\beta}$ (shared by all N frames) and pose $\boldsymbol{\theta}_n$ (specific to the n-th frame), resulting in a $26N + 112$ dimensional optimization. This non-linear and non-convex problem is solved in a Levenberg fashion though iterative linearization from a given initialization $\{\boldsymbol{\Psi}_n^{t=0}\}$:

$$\underset{\{\Delta\boldsymbol{\Psi}_n\}}{\arg\min} \sum_{n=1}^{N} \sum_{\tau \in \tau_{\text{calib}}} \omega_\tau E_\tau(\Delta\boldsymbol{\Psi}_n; \mathcal{D}_n) \tag{2}$$

$$\boldsymbol{\Psi}_n^{t+1} = \boldsymbol{\Psi}_n^{t} + \Delta\boldsymbol{\Psi}_n \tag{3}$$

We use the following four calibration energy terms τ_{calib}:

d2m	each data point is explained by the model
m2d	the model lies in the sensor visual-hull
valid	prevent invalid/degenerate sphere-meshes
reg	prioritize regularization of shape vs. pose

This formulation extends the framework in Tkach et al. [28], and could be rewritten as MAP optimization with Laplacian distributions for the fitting terms, and Gaussian distributions for regularization priors. We refer the reader to [28] for a detailed description of {*d2m*,*m2d*,*valid*}. Notably, our optimization differs in two ways from the one proposed in [28]: (1) we introduce an *update-regularization*, (2) we can omit the *rigidity* prior. These changes profoundly impact the robustness of calibration optimization, but require the *re-parameterization* of our calibration DOFs. In contrast to [28], our re-parameterization decouples pose parameters $\boldsymbol{\theta}$ (e.g. finger bend angle) from shape parameters $\boldsymbol{\beta}$ (e.g. phalanx length); see Section 3.1.

Initialization and update-regularization. We employ the default template and track in real-time the movements of the user, thus generating our initialization $\{\boldsymbol{\Psi}_n^{t=0}\}$. Then, we penalize large updates to prevent large linearization errors

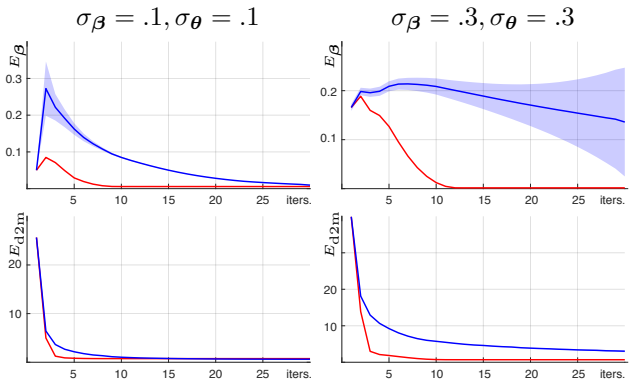

Figure 5: In this test, we calibrate a single finger and show how the convergence properties improve switching from a **position-based** [28] to our **decoupled** $[\boldsymbol{\theta}_n, \boldsymbol{\beta}]$ formulation. The shading denotes one standard deviation σ of uncertainty; note that although variance is always plotted, at times it is not visible (i.e. $\sigma \approx 0$). $E_{\boldsymbol{\beta}}$ and E_{d2m} are respectively defined in Eq. 13 and [28].

by employing a Tikhonov regularizer:

$$E_{\mathrm{reg}}(\Delta\boldsymbol{\Psi}_n; \mathcal{D}_n) = \|\Delta\boldsymbol{\Psi}_n\|^2_{\mathbf{H}} = \Delta\boldsymbol{\Psi}_n^T \, \mathbf{H} \, \Delta\boldsymbol{\Psi}_n \qquad (4)$$

However, as our initialization consists of *posed* models, we legitimately assume that pose parameters are relatively close to desired ones; that is, $\{\boldsymbol{\theta}_n^{t=0}\} \approx \{\boldsymbol{\theta}_n^*\}$. The diagonal matrix $\mathbf{H}$ allows us to encode such requirement, where $\mathbf{H}_{kk} = \omega_{\text{pose}}$ if the k-th entry is a *pose*-DOF and $\mathbf{H}_{kk} = \omega_{\text{shape}}$ if the k-th entry is a *shape*-DOF, and we select $\omega_{\text{pose}} \gg \omega_{\text{shape}}$. We evaluate the impact of such differential treatment in Figure 4. When calibration is executed on sensor data, where self-occlusions and structured outliers result in incomplete depth maps, we noted damping pose updates further stabilized our optimization avoiding the algorithm from falling into local minima.

Reduced-DOF optimization (shape-space). Recent results by Tan et al. [24] indicate that solving for calibration in a low-dimensional shape-space tends to widen the basin of convergence of the optimization. A simple shape-space controlled by a single degree of freedom can also be constructed by considering a calibration that uniformly rescales the default template; this simple adjustment has been shown to have a considerable impact on tracking accuracy [28]. We now generalize this concept, and introduce a shape-space that represents *locally-anisotropic scaled* versions of the default template model. The definition of such a subspace, differently from [24], does not require the collection and processing of large amounts of manually annotated data, and its DOFs are easy to interpret (e.g. finger lengths and girth, palm width and height, ...). Given the default shape parameters $\bar{\boldsymbol{\beta}}$ and a perturbation vector $\tilde{\boldsymbol{\beta}}$, any (anisotropic) deformation of the template with default shape parameters $\bar{\boldsymbol{\beta}}$ can be obtained through the transformation:

$$\boldsymbol{\beta} = \bar{\boldsymbol{\beta}} \circ [\mathbf{1} + \tilde{\boldsymbol{\beta}}] \qquad (5)$$

where $\circ$ represents the element-wise product. We can now obtain a low-dimensional shape-space by the simple process of *clustering* degrees of freedom into groups; for example, if all finger lengths belong to the same group they will all be scaled by the same $\tilde{\beta}_{\text{finger length}}$. As we detail in Section 3.2, we propose a 5 dimensional shape-space manifold whose DOFs represent {palm aspect-ratios, sphere radii and finger lengths}, and demonstrate how anisotropic deformations such as those described in Eq. 5 can be easily embedded in Levenberg calibration frameworks. The shape-space reduces the number of local minima, widens the basin of convergence, and makes the calibration algorithm more robust: having it converge to the correct solution even when starting far away from the global minimum; see Figure 8.

Multi-stage optimization. However, reducing the number of degrees of freedom with a shape-space overly constraints the template; see Figure 3b. To resolve this issue, we propose a two-stages optimization; see Figure 3c. First, a coarse-scale robust optimization is executed in a *reduced* 5 DOF space (**stage 1**), followed by a calibration in the *full* 112 DOF space to refine the model (**stage 2**). The multi-stage calibration performance is reported in Figure 9.

3.1. Parameterizing the calibration energy

The adoption of the effective shape-space in Equation 5 requires a re-parameterization of the degrees of freedom controlling the shape of our tracking template. In this section, we first describe the parameterization from Tkach et al. [28] and highlight its shortcomings. Then we introduce our novel parameterization, whose improved performance is evaluated in Section 4 and illustrated in Figure 5.

Position-based parameterization. The sphere-mesh model of Tkach et al. [28] employs two matrices with sphere positions and corresponding radii $(\bar{\boldsymbol{C}}, \bar{\boldsymbol{r}})$ to describe rest-pose geometry of the hand model. A kinematic chain with rest-transformations $\{\bar{\boldsymbol{T}}_k\}$ deforms the sphere-mesh through pose parameters $\boldsymbol{\theta}$, producing a posed model $\boldsymbol{C}_n$ to fit the data in frame $\mathcal{D}_n$. Therefore, the *shape* degrees of freedom of such a model are $\bar{\boldsymbol{C}}, \bar{\boldsymbol{r}}, \{\bar{\boldsymbol{T}}_k\}$ calibrated by [28] through a two-step optimization with terms $\tilde{\tau}_{\text{calib}} \neq \tau_{\text{calib}}$:

$$\begin{aligned} &\underset{\{\boldsymbol{C}_n\},(\bar{\boldsymbol{C}},\bar{\boldsymbol{r}})}{\arg\min} \sum_{n=1}^{N} \sum_{\mathcal{T}\in\tilde{\tau}_{\text{calib}}} w_{\mathcal{T}} E_{\mathcal{T}}(\mathcal{D}_n, \boldsymbol{C}_n, \bar{\boldsymbol{C}}, \bar{\boldsymbol{r}}) \\ &\underset{\{\boldsymbol{\theta}_n\},\{\bar{\boldsymbol{T}}_k\}}{\arg\min} \sum_{n=1}^{N} \sum_{\mathcal{T}\in\tilde{\tau}_{\text{calib}}} w_{\mathcal{T}} E_{\mathcal{T}}(\boldsymbol{C}_n, \boldsymbol{\Psi}_n) \end{aligned} \qquad (6)$$

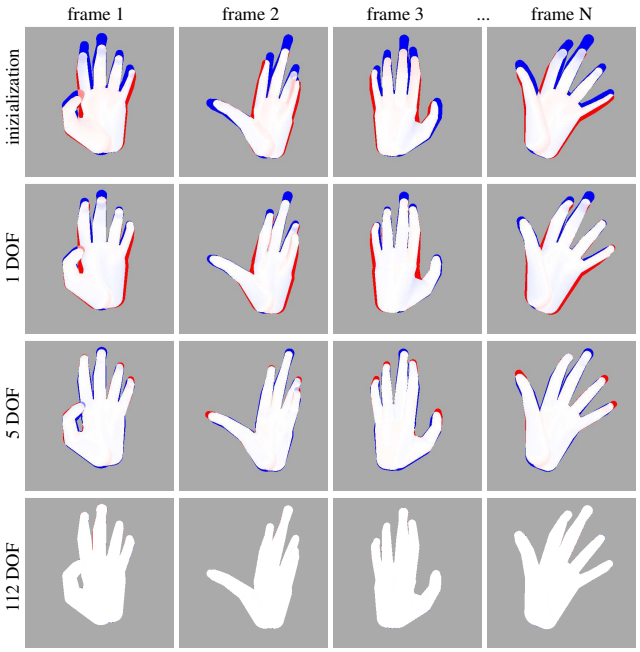

Figure 6: Our parameterization allows the construction of a 1 DOF (uniform scaling) and a 5 DOF shape-space (anisotropic palm plus finger length/thickness). The full-DOF shape-space (last row) can recover a fully calibrated model but only when a sufficiently good initialization, such as the one visualized here, is provided.

In [28], a two-step process is necessary as the posed sphere-meshes $\{\boldsymbol{C}_n\}$ *entangle* pose and shape parameters; that is, $\boldsymbol{\theta}_n = f_1(\boldsymbol{C}_n)$ and $\{\bar{\boldsymbol{T}}_k\} = f_2(\{\boldsymbol{C}_n\})$ are represented by $\{\boldsymbol{C}_n\}$ through non-linear mappings f_1, f_2. In particular, note how the optimization in Eq. 6 addressed this issue by *de-coupling* these degrees of freedom through an alternating optimization: first by optimizing over posed-shapes, then optimizing over shape parameters and deducing appropriate joint angles. Our experiments show how this entanglement can be detrimental, resulting in poor convergence properties for the algorithm in Eq. 6; see Section 4.

Decoupled shape-pose parameterization. Our work stems from a fundamental assumption: the rotational part of kinematic frames in the rest pose $\{\bar{\boldsymbol{T}}_k\}$ is shared by all users. This is a justified choice in our context by two observations: (1) orthopaedic research by Hollister et al. [6] has revealed this assumption is valid for typical human hands, and (2) our default template, inherited from [28], contains optimized kinematic frames $\{\bar{\boldsymbol{T}}_k\}$; see Equation 6. Further, rather than representing each finger's geometry as tuples of centers/radii $\{(\boldsymbol{c}_i, r_i)\}$ (16 DOF) as in [28], we encode finger-palm attachment $\boldsymbol{c}_i$ and the length δ_i and girth r_i of each phalanx. The pose of each finger is then expressed by a 4-tuple of angles $\{\theta_1^a, \theta_1^f, \theta_2^f, \theta_3^f\}$ (14 DOF),

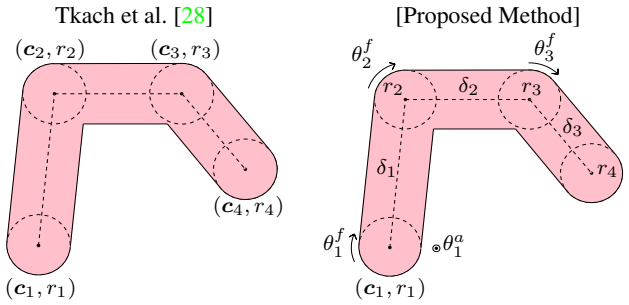

Figure 7: Sphere-mesh vs. our $[\boldsymbol{\theta}, \boldsymbol{\beta}]$ parameterization.

where a stands for *abduction* and f represents *flexion*. The parameterization of centers coordinates for palm, wrist and flexible sphere-mesh elements are unchanged, leading to a 112 DOF model. To reflect the decoupling of pose and shape, we denote our parameter vector as $\boldsymbol{\Psi}_n = [\boldsymbol{\theta}_n, \bar{\boldsymbol{\beta}}]$, where $\bar{\boldsymbol{\beta}} = [\bar{\delta}, \bar{\boldsymbol{r}}, \{\bar{\boldsymbol{c}}\}_{\text{palm}}]$.

3.2. Shape-space optimization

Considering the anisotropic perturbation model of Eq. 5 and recalling the parametrization described in Section 3.1, our shape-manifold is obtained by coalescing the following degrees of freedom:

$$\begin{aligned}\boldsymbol{\delta} &= \bar{\boldsymbol{\delta}}(1 + \tilde{\delta}),\\ \boldsymbol{r} &= \bar{\boldsymbol{r}}(1 + \tilde{r})\\ \boldsymbol{c} &= \bar{\boldsymbol{c}} \circ ([1,1,1]^\intercal + \tilde{\boldsymbol{c}})\end{aligned}$$

That is, inspired by Khamis et al. [7], we employ a 5 dimensional shape-space, whose degrees of freedom are selected by inspection of its main modes of variation [7, Fig.1]. The shape-space, parameterized by $\tilde{\boldsymbol{\beta}} = [\tilde{\delta}, \tilde{r}, \tilde{\boldsymbol{c}}]$, encodes:

$$\begin{aligned}\tilde{\delta}&: \quad \text{uniform finger-length scaling}\\ \tilde{r}&: \quad \text{uniform sphere-scaling}\\ \tilde{\boldsymbol{c}}&: \quad \text{anisotropic 3D palm stretching}\end{aligned}$$

One could re-differentiate the calibration energies in Eq. 2 with respect to these new degrees of freedom – a tedious and laborious task. However, recall how the energy terms in Eq. 1 can be written in squared-residual form. Without loss of generality, let us narrow our consideration to $\boldsymbol{\delta}$, that is, the DOF representing finger lengths:

$$\arg\min_{\boldsymbol{\delta}^{t+1}} \|\boldsymbol{\epsilon}(\boldsymbol{\delta}^{t+1})\|^2 \tag{7}$$

To derive the Levenberg update, we perform the first order Taylor expansion of Eq. 7 with respect to $\boldsymbol{\delta}$:

$$\arg\min_{\Delta\boldsymbol{\delta}^{t+1}} \|\boldsymbol{\epsilon}(\boldsymbol{\delta}^{t}) + J_{\boldsymbol{\delta}}\Delta\boldsymbol{\delta}^{t+1}\|^2 \tag{8}$$

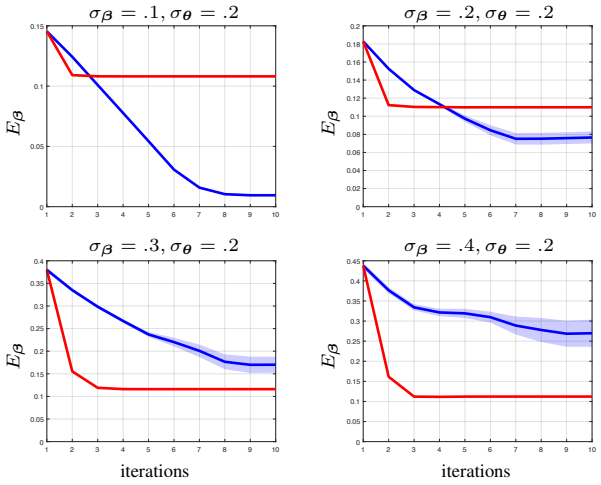

Figure 8: A comparison of the convergence properties of the re-parameterized **full-DOF** (112 parameters) vs. **reduced-DOF** (5 parameters) calibration. As we increase the perturbation, the full-DOF optimization loses robustness as illustrated by the variance plots. Conversely, the reduced-DOF optimization is highly robust, although it only provides a rough approximation of the parameters.

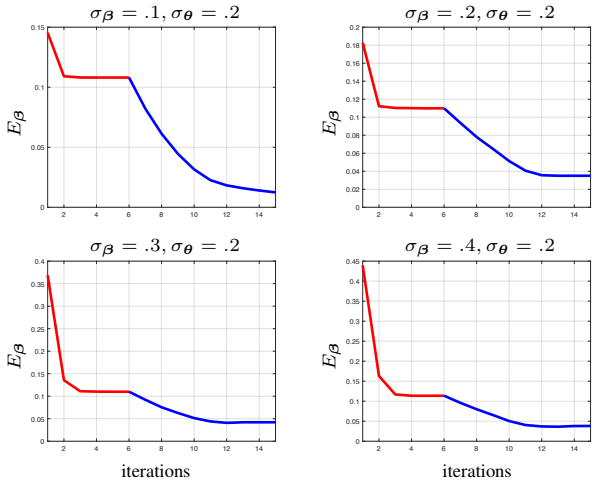

Figure 9: The 2-stage algorithm outlined in Figure 3 combines the robustness of the **reduced-DOF** to the accuracy of the **full-DOF** calibration. For visualization purposes only, we execute five iterations of reduced-DOF optimization; note in all our experiments, five iterations are sufficient to achieve convergence of the low-dimensional solve. Note how variance ≈ 0, revealing enhanced convexity.

where $J_{\boldsymbol{\delta}} = \partial \boldsymbol{\epsilon} / \partial \boldsymbol{\delta}$ is the Jacobian matrix. Leveraging our shape-space, we rewrite the differential update as:

$$\boldsymbol{\delta}^{t+1} = \boldsymbol{\delta}^t + \Delta\boldsymbol{\delta}^{t+1} = \boldsymbol{\delta}^t(1 + \tilde{\delta}^{t+1}) \;\Rightarrow\; \Delta\boldsymbol{\delta}^{t+1} = \boldsymbol{\delta}^t \tilde{\delta}^{t+1}$$

effectively re-parameterizing our optimization:

$$\underset{\tilde{\delta}^{t+1}}{\arg\min} \|\boldsymbol{\epsilon}(\boldsymbol{\delta}^t) + (J_{\boldsymbol{\beta}}\boldsymbol{\delta}^t)\tilde{\delta}^{t+1}\|^2 \tag{9}$$

Differentiating the energy above with respect to the unknown $\tilde{\delta}^{t+1}$ leads to the least-squares solution:

$$\tilde{\delta}^{t+1} = (J_{\boldsymbol{\delta}}\boldsymbol{\delta}^t)^\dagger \boldsymbol{\epsilon}(\boldsymbol{\delta}^t) \;\Rightarrow\; J_{\tilde{\delta}} = J_{\boldsymbol{\delta}}\boldsymbol{\delta}^t \tag{10}$$

In summary, the Jacobians computed for an existing full-DOF optimization can be *recycled* to optimize in a desired low-dimensional shape-space. This allows the proposed anisotropic scaling shape-space to be embedded in any Levenberg calibration algorithm with minimal changes to the source code.

4. Evaluation

We evaluate our algorithm on synthetic data as well as datasets acquired from a commodity RGBD sensor. Our calibration algorithm is developed in MATLAB and executed single-threaded on a 3GHz laptop. The average execution time for $N = 8$ is 200s, and its complexity grows linearly in N as well as in the depth image resolution. As the optimization problem is analogous in nature to [23], a GPU or C++ implementation can be expected to achieve real-time performance (respectively for dense or sub-sampled data; see [27] and [29]). In the evaluation that follows, we study the robustness of our algorithm by analyzing its convergence properties on synthetic data. We also evaluate our calibration on raw sensor data, confirming that calibrated sphere-meshes [28] can achieve competitive tracking precision compared to triangular meshes [27].

Evaluation on synthetic data. We quantitatively evaluate the proposed system by randomly perturbing a ground truth model and studying the convergence properties of the algorithm (convergence speed and robustness). Ground-truth parameters $[\boldsymbol{\beta}^*, \{\boldsymbol{\theta}_i^*\}]$, where i indexes a given pose, are perturbed to generate $M = 100$ initializations $\{[\boldsymbol{\beta}^{t=0}, \{\boldsymbol{\theta}_i^{t=0}\}]_{m=1\ldots M}\}$ with uniform distribution:

$$\boldsymbol{\beta}^{\text{t=0}} = \boldsymbol{\beta}^* \circ [\mathbf{1} + \mathcal{U}(\mathbf{0}, \sigma_{\boldsymbol{\beta}})] \tag{11}$$

$$\boldsymbol{\theta}_i^{\text{t=0}} = \boldsymbol{\theta}_i^* + \mathcal{U}(\mathbf{0}, \sigma_{\boldsymbol{\theta}}) \tag{12}$$

We evaluate robustness by measuring the ability of the algorithm to return to the optimal configuration $[\boldsymbol{\theta}^*, \boldsymbol{\beta}^*]$ over a randomly sampled set of perturbations as we vary the magnitudes of $\sigma_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\beta}}$ as measured by the error metric:

$$E_{\boldsymbol{\beta}} = \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \tag{13}$$

In Figure 5, we investigate the effect of re-parameterization. Here, we simplify the problem by considering a single finger in isolation. Note how decoupling pose and shape results in substantially improved convergence properties. In

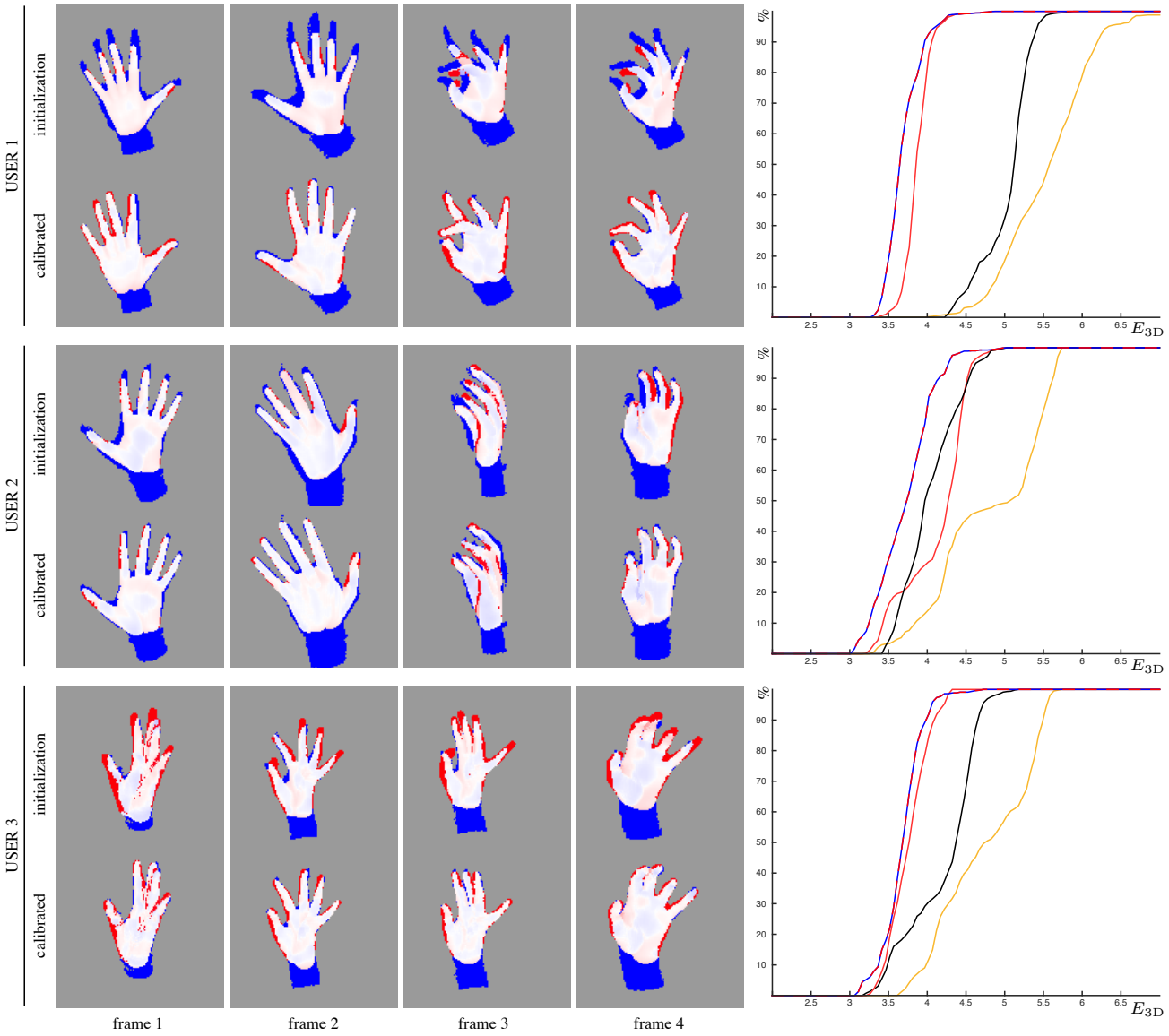

Figure 10: Model personalization evaluation of captured depth data for different users. We evaluate tracking performance as the percentage of frames where the E_{3D} metric [28] is lower than the value in the abscissa for the **template**, **reduced-DOF** (i.e. shape-space), and our **multi-stage** calibrated model. We do not report the curve for neither the calibration of [28], nor for our **full-DOF**, as they could not converge to a consistent solution, revealing their reliance on a good initialization. We also evaluate against the state-of-the-art tracking algorithm by Taylor et al. [27], whose calibration was performed via Tan et al. [24]; the rendered depth-maps to evaluate the metrics are courtesy of the authors. These results are better appreciated in the videos available on the project page.

Figure 8, we consider full hand calibration and we gradually increase the initialization perturbation and thus the difficulty of the problem. The full-DOF optimization has difficulties in recovering the ground-truth parameters when the perturbation is too large; conversely, the shape-space calibration optimization is robust, but fails at estimating precise parameter values. A qualitative example of this shortcoming is presented in Figure 6, where we illustrate how only the full-DOF parameterization can produce a sufficiently accurate model. As illustrated in Figure 9, our *multi-stage* calibration combines the strengths of the two approaches by first robustly optimizing the rough template shape in a low-dimensional subspace, and then refining the model in the high-dimensional space.

Evaluation on acquired data. We acquired tracking/calibration sequences for different users using an *Intel RealSense SR300* depth sensor. Tracking is performed with the real-time algorithm by Tkach et al. [28]. Note that while tracking the shape parameters $\boldsymbol{\beta}$ are kept fixed. Tracking quality is quantitatively evaluated with the *algorithmic agnostic* metric E_{3D} described in [28]. A few calibration frames and an illustration of the model alignment before/after calibration optimization is executed are reported in Figure 10. In this figure we also report the E_{3D} tracking metric for the template vs. the personalized model. We do not report the curve for [28], as the algorithm fails to converge with poor initializations such as the ones we use. In Figure 11, we compare a number of state-of-the-art techniques on the complex HANDY/TEASER dataset from [28]. Note how our calibration leads to a slight loss in tracking performance when compared to [28], but is still competitive to the tracking performance of Taylor et al. [27], that employs a model calibrated by the algorithms proposed by Tan et al. [24]. The tracking performance of [28] is still slightly superior to the one of our method. This is expected, as Tkach et al. [28] calibrates on curated *panoptic* input data (a dense point cloud computed via multi-view stereo (MVS) on hundreds of RGB images, generating multiple poses and ≈4k points/pose), and manual annotation was used to guide the algorithm. Instead, our solution leverages a small set of low-resolution depth maps (typically 8), and requires no manual intervention.

5. Conclusions and future work

In this paper, we presented a novel sphere-mesh calibration algorithm leveraging a low dimensional shape-space to robustly personalize a default tracking template to a given user. We qualitatively and quantitatively evaluated the algorithm on several users against the state-of-the-art calibration techniques of Tkach et al. [28] and Tan et al. [24], revealing how our calibrated models achieve competitive tracking performance. In comparison to [28], we demonstrated our algorithm possesses a much wider basin of convergence, simplifying the calibration task by removing the need for initialization fine-tuning. Conversely from Tan et al. [24], which relies on differentiation of an expensive render-and-compare energy, our calibration framework is *analytically differentiable* with respect to pose/shape parameters. Given the signal-to-noise ratio of current generation real-time depth sensors, our investigation also reveals how simple geometric observations can be used to build a shape-space that is effective in regularizing calibration optimization. These characteristic allowed us to extend the approach we presented in this paper into [29], hence introduce the first online/streaming calibration technique for hand tracking. Further, the technique of Tan et al.

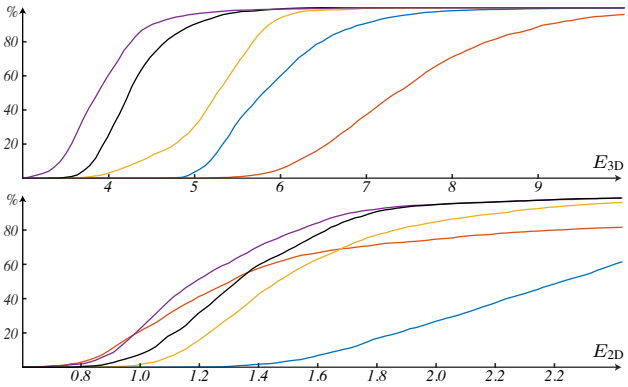

Figure 11: The tracking performance of the model calibrated by the **[Proposed Method]** is quantitatively evaluated against the ones of **Tkach et al. [28]**, **Tagliasacchi et al. [23]**, **Sharp et al. [18]** and **Taylor et al. [27]** on the HANDY/TEASER sequence. A minor loss in performance is expected, as [28] calibrated this dataset on high-quality MVS data, while we employ consumer-level depth imagery.

[24] is model-specific, and requires some form of ad-hoc retargeting in order to be adopted in other tracking systems; conversely, our shape-space does not require to repeat the acquisition, calibration and analysis of a large set of users, but can be easily deployed with minimal intervention in existing tracking codebases. In the future, we intend to investigate whether a data-driven shape-space can be used to represent *fine-scale* geometry differences, and potentially account for pose-space deformations [3]. We leave to future work to determine whether a multi-stage algorithm that increases the DOFs more gradually would be advantageous; such coarse-to-fine schemes are already known to be beneficial for hand tracking [25]. Through our shape-space, we can deal with deformation in a $\pm 40\%$ range, thus covering a large portion of the human population. The range could be further extended by requesting user to sign-in into the calibration session with a default rest-pose from which a uniform scale can roughly be estimated. In a similar direction, a set of standard calibration poses could be used, and pose latent spaces [23] that are *pose-specific* could be used as regularizers in place of our pose-update damping strategy. To ensure reproducibility of results, as well as to foster further research in this area, we release the **sources** of our algorithm as well as all our **evaluation datasets**. These resources are available at the page: `http://github.com/edoRemelli/hadjust`.

Acknowledgments. We are grateful to L. Pegolotti and S. Shoulder for providing the data used in these experiments. This research is supported by the SNF grant #200021-153567, the NSERC Discovery grant #2016-05786 and the Google/Intel Industrial Research Chair in 3D Sensing.

References

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

[2] L. Ballan, A. Taneja, J. Gall, L. V. Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In *Proc. ECCV*, 2012.

[3] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proc. ICCV*, 2015.

[4] M. de La Gorce, D. J. Fleet, and N. Paragios. Model-based 3D hand pose estimation from monocular video. *PAMI*, 2011.

[5] L. Dipietro, A. Sabatini, and P. Dario. A survey of glove-based systems and their applications. *IEEE Transactions on Systems, Man, and Cybernetics*, 38(4), 2008.

[6] A. Hollister, W. L. Buford, L. M. Myers, D. J. Giurintano, and A. Novick. The axes of rotation of the thumb carpometacarpal joint. *Journal of Orthopaedic Research*, 1992.

[7] S. Khamis, J. Taylor, J. Shotton, C. Keskin, S. Izadi, and A. Fitzgibbon. Learning an efficient model of hand shape variation from depth images. *Proc. CVPR*, 2015.

[8] D. Kim, O. Hilliges, S. Izadi, A. D. Butler, J. Chen, I. Oikonomidis, and P. Olivier. Digits: freehand 3d interactions anywhere using a wrist-worn gloveless sensor. *Proc. ACM UIST*, 2012.

[9] A. Makris and A. Argyros. Model-based 3d hand tracking with on-line hand shape adaptation. *Proc. BMVC*, 2015.

[10] S. Melax, L. Keselman, and S. Orsten. Dynamics based 3d skeletal hand tracking. In *Proc. of Graphics Interface*, 2013.

[11] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a Feedback Loop for Hand Pose Estimation. In *Proc. ICCV*, 2015.

[12] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. In *Proc. Computer Vision Winter Workshop*, 2015.

[13] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3D tracking of hand articulation using kinect. In *Proc. BMVC*, 2011.

[14] G. Pons-Moll and B. Rosenhahn. *Model-Based Pose Estimation*. Springer, 2011.

[15] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *Proc. CVPR*, 2014.

[16] J. M. Rehg and T. Kanade. Visual tracking of high dof articulated structures: An application to human hand tracking. In *Proc. ECCV*, 1994.

[17] T. Rhee, U. Neumann, and J. P. Lewis. Human hand modeling from surface anatomy. In *Proc. ACM i3D*, 2006.

[18] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proc. of ACM CHI*, 2015.

[19] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using RGB and depth data. In *Proc. ICCV*, 2013.

[20] S. Sridhar, H. Rhodin, H.-P. Seidel, A. Oulasvirta, and C. Theobalt. Real-time hand tracking using a sum of anisotropic gaussians model. In *Proc. International Conference on 3D Vision (3DV)*, 2014.

[21] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proc. CVPR*, 2015.

[22] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *Proc. ICCV*, 2015.

[23] A. Tagliasacchi, M. Schroeder, A. Tkach, S. Bouaziz, M. Botsch, and M. Pauly. Robust articulated-icp for real-time hand tracking. *Computer Graphics Forum (Proc. SGP)*, 2015.

[24] D. J. Tan, T. Cashman, J. Taylor, A. Fitzgibbon, D. Tarlow, S. Khamis, S. Izadi, and J. Shotton. Fits like a glove: Rapid and reliable hand shape personalization. In *Proc. CVPR*, 2016.

[25] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton. Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In *Proc. ICCV*, 2015.

[26] 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 *Proc. CVPR*, 2014.

[27] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, J. Valentin, B. Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 2016.

[28] A. Tkach, M. Pauly, and A. Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2016.

[29] A. Tkach, A. Tagliasacchi, E. Remelli, M. Pauly, and A. Fitzgibbon. Online generative model personalization for hand tracking. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2017.

[30] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 2014.

[31] T. von Marcard, B. Rosenhahn, M. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum (Proc. of Eurographics)*, 2017.

[32] R. Y. Wang and J. Popovic. Real time hand tracking with a colored glove. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 2009.

[33] G. Welch and E. Foxlin. Motion tracking survey. *IEEE Computer Graphics and Applications*, 2002.

[34] R. Xu, S. Zhou, and W. J. Li. Mems accelerometer based nonspecific-user hand gesture recognition. *IEEE sensors journal*, 2012.

[35] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T. Kim. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. 2017.

[36] W. Zhao, J. Chai, and Y.-Q. Xu. Combining marker based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *Proc. SCA*, 2012.