Automatic Bone Parameter Estimation for Skeleton Tracking in Optical Motion Capture

Tobias Schubert Katharina Eggensperger Alexis Gkogkidis
Frank Hutter Tonio Ball Wolfram Burgard

***Abstract*— Motion analysis is important in a broad range of contexts, including animation, bio-mechanics, robotics and experiments investigating animal behavior. For applications, in which tracking accuracy is one of the main requirements, passive optical motion capture systems are widely used. Many skeleton tracking methods based on such systems use a predefined skeleton model, which is scaled once in the initialization step to the individual size of the character to be tracked. However, there are remarkable differences in the bone length relations across gender and even more across mammal races. In practice, the optimal skeleton model has to be determined in a manual and time-consuming process. In this paper, we reformulate this task as an optimization problem aiming to rescale a rough hierarchical skeleton structure to optimize probabilistic skeleton tracking performance. We solve this optimization problem by means of state-of-the-art black-box optimization methods based on sequential model-based Bayesian optimization (SMBO). We compare different SMBO methods on three real-world datasets with an animal and humans, demonstrating that we can automatically find skeleton structures for previously unseen mammals. The same methods also allow an automated choice of a suitable starting frame for initializing tracking.**

I. Introduction

Motion capture systems are widely used to detect human and animal postures in motion. Typical applications include animation, bio-mechanics, robotics and experiments investigating animal behavior. There exist different approaches to detect the skeleton structure from information about the exterior of the mammal to be tracked. Depending on the application at hand, one can use inertial measurement units, passive markers, which are attached to a human or animal, or other approaches that use RGB-D data to extract motion.

For experiments investigating animal behavior, accuracy is one of the principal requirements since one usually needs to detect small changes in the movements (for example due to training, medication, or brain stimulation). Compared to marker-less approaches, optical motion capture systems based on passive markers yield smoother motion trajectories at high frame rates [22] and are normally more robust against occlusions. In contrast to systems based on inertial measurement units, passive marker based systems do not impede the animal in motion. However, a major drawback of these systems is that they return frames of unlabeled marker positions (see, e.g., the top left image in Fig. 1).

All authors are with the University of Freiburg, Germany. This work has partly been supported by the German Research Foundation under grant EXC 1086. The recording of the sheep datasets was approved by the administrative council in Freiburg and we adhere to the Protection of Animals Act and the "Tierschutz-Versuchstierordnung".

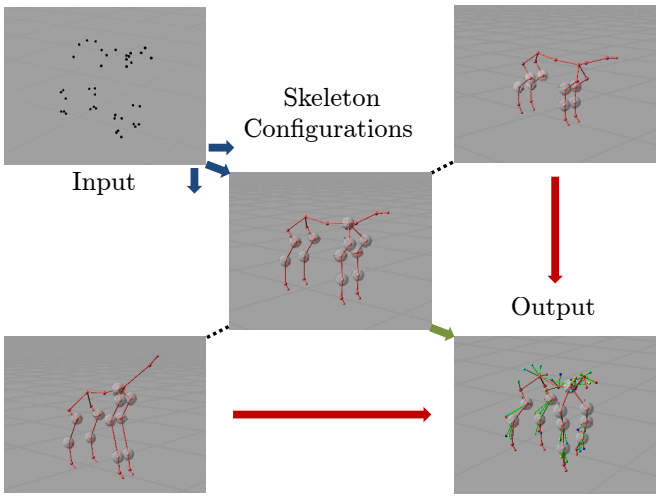


Fig. 1: Top left: Unlabeled marker positions given by the Motion Analysis software Cortex [10]. Diagonal bottom left to top right: Possible skeleton configurations. Bottom right: Fit of the best skeleton configuration into the marker-cloud.

This introduces the problem of identifying the best of several possible skeleton configurations that best explain these marker positions (see the remainder of Fig. 1).

For an automatic labeling procedure, the occlusion of markers over longer periods of time and the labeling of reappearing markers are challenging problems. Furthermore, the markers are attached to the skin or coat of the tracked animal or human and therefore move during motion. Several methods exist for addressing these problems. Meyer et al. [20] presented a skeleton tracking method that allows to place markers at arbitrary positions onto the human and automatically labels them between frames. Schubert et al. [25] extended this method with an automatic initialization method for many kinds of initial poses. However, in order to apply these methods, one has to manually define the relative bone length, muscle width, and initial frame, all of which influence the accuracy.

Unfortunately, measuring the bone lengths of animals often requires to narcotize them and in addition proved to be inaccurate. In practice, we obtained the best tracking results by manually tuning the bone parameters while taking the tracking performance into account. In the following, we call the skeleton we obtained by this iterative manual trial-and-error procedure "ground truth skeleton". This time-consuming manual process needs to be reiterated for each

individual, for each gender and even more so for other mammals races. Alternative solutions, which are able to compute the skeleton structure, require manually labeled marker frames or a specific minimal number of markers attached to each modeled segment. So far, no system exists that can deal with a limited number of markers and can initialize the tracker without manual labeling or measurements of bone lengths and muscle widths.

In this paper we present a method that automatically optimizes these bone parameters. To enable this optimization, we define an appropriate performance measure for skeleton tracking that encompasses several aspects of a reasonable tracking, including the percentage of missed markers, the goodness of the skeleton structure fit into the marker cloud, and the difference between the predicted and true marker positions. We optimize this performance measure with recent state-of-the-art global optimization techniques based on sequential model-based Bayesian optimization [17], which automatically finds skeletons that minimize the tracking error without manual intervention.

We apply our method to the skeleton tracking of a sheep (which is increasingly used as a large animal model in biological studies[1]), a female human and a male human. Our empirical results demonstrate that our method can automatically find a skeleton structure close to the ground truth skeleton, which leads to smoother motion trajectories (Fig. 3). We also show, that we can extend our approach to find an optimal start frame for initializing the skeleton tracking (Fig. 4).

II. Related Work

In the last few years marker-less approaches for skeleton tracking gained popularity, see e.g. Moeslund et al. [21], Elhayek et al. [14], Baak et al. [2], Yan and Pollefeys [31]. Several methods infer the skeleton structure out of dense point cloud data, see e.g. Au et al. [1], Huang et al. [16]. Unfortunately, the accuracy of marker-less approaches is not high enough for certain applications, such as medical studies, where one wants to detect small movement changes [15, 26]. Therefore, we consider a marker-based system for tracking.

One sub-task of a skeleton tracking method is the computation of the underlying basic skeleton structure. One approach is to detect the regions of rigid body segments first by examination of relative marker distances and to infer the position of the joints connecting these segments, see Ringer and Lasenby [23], Kirk et al. [18], and de Aguiar et al. [12]. These methods need a manual labeling step (which is time-consuming and tedious work) and require a certain number of markers associated to each segment. Furthermore, the computation of the skeleton structure is done during tracking and thus slows down the attainable tracking frame rate. In the constrained scenarios of medical studies or animal experiments one can assume, that the subject to be tracked is the same during an experimental study and the skeleton bone lengths and muscle widths can be computed beforehand. Besides, a method which robustly works for different marker placements and counts is desirable.

There are extensive studies for the skeleton bone lengths and muscle widths of humans [9], but for most of the other mammals the underlying skeleton structure is only known approximately. Differences due to gender and mammal races (Contini [9] showed that there are even remarkable differences between male and female human skeleton structures) are not taken into account.

The work presented in this paper add an automatic determination of bone lengths and muscle widths to the work of Meyer et al. [20] and Schubert et al. [25], by the introduction of a more reasoned skeleton tracking objective function.

III. Problem Formulation

Our input data is a set $F_{1:T}$ of frames of unlabeled three-dimensional observations $z_{i,t} \in F_t$ at equidistant discrete time steps t. Each observation $z_{i,t}$ is the three-dimensional position of one visible marker $m \in M$ attached to the skin, cloth, or coat of the object to be tracked. First of all we have to find the skeleton bone parameter vector B, containing the bone lengths and muscle widths. For a given skeleton bone parameter, the set of reachable skeleton configurations $\mathcal{C}(B)$ is then defined by varying the translation of the root segment and the rotations of each segment of the skeleton model. The aim of the skeleton tracking is then to infer the skeleton configuration $C_t \in \mathcal{C}(B)$ for each time step t.

A. Skeleton Model

We use a predefined skeleton model, which consists of a tree structure of joints and segments, with varying bone lengths and muscle widths. For example, in the case of a sheep it consists of 25 segments. Due to the sparse coverage of the body with markers, one cannot reasonably compute the length of the outermost limbs, and thus this number can be reduced to 20 segments. Although it is not feasible to model each vertebra in the backbone, it suffices to give a good approximation of the problem, while keeping the optimization dimension low. The skeleton model is hierarchical as in Meyer et al. [20], namely for given bone parameter vector B each skeleton configuration in the set of reachable skeleton configurations $\mathcal{C}(B)$ can be described uniquely by a three-dimensional rotation for each segment plus a three-dimensional translation vector of the root segment. All together, we describe the skeleton configuration of a sheep by $57 = 3+3\cdot 18$ degrees of freedom (see Fig. 1), where the 18 comes from the fact that we model the root segment with three bones but only one orientation, where we represent the rotations using unit quaternions. In the case of a sheep the bone parameter vector B is 40-dimensional, which can be reduced to a 24-dimensional vector (12 dimensions of bone lengths and muscle widths, respectively), if we take the symmetry between left and right into account.

B. Probabilistic Skeleton Tracking

There is a lot of uncertainty hidden in the problem formulation, which we have to take into account. The observed

[1]While we use a sheep as the only animal in our experiments, our method can be easily extended to other kinds of mammals.

positions $F_{1:T}$ are affected by measurement noise. The markers are attached to the skin, cloth or coat of the object and thus move slightly and non-deterministically with respect to the corresponding segment. This suggests a probabilistic problem formulation. Assume first, that the bone parameter vector B is given. The goal of a skeleton tracking algorithm is then to find the most likely skeleton configurations $C^*_{1:T}$ given the marker observations $F_{1:T}$. Formally, this means that we have to solve

$$C^*_{1:T}(B) = \arg\max_{C_{1:T}} P\left(C_{1:T}(B) \mid F_{1:T}\right), \qquad (1)$$

for a given B. In the work of Schubert et al. [25] and of Meyer et al. [20] this equation is approximated to allow an online skeleton tracking. Assume we can compute, or at least approximate, the most likely skeleton configurations $C^*_{1:T}(B)$ for a given bone parameter vector B. The goal is then to find the bone parameter vector B^* for which the skeleton configuration $C^*_{1:T}(B)$ describes the marker observations $F_{1:T}$ best. This means we have to solve

$$B^* = \arg\max_{B} P\left(C^*_{1:T}(B) \mid F_{1:T}\right). \qquad (2)$$

In the work of Schubert et al. [25] and Meyer et al. [20] the most likely skeleton configuration $C^*_t(B)$ is computed iteratively from the previous one $C^*_{t-1}(B)$. They first use the Hungarian method [19] to associate the unlabeled marker observations $\{z_{i,t}\}$ to markers. Given the skeleton configuration C_t, they use the relative positions R_{t-1} of markers to the corresponding bones to predict the global position $p_{i,t}(C_t, R_{t-1})$ of marker i, which is associated to observation $z_{i,t}$ at time t. Finally, they obtain the most likely skeleton configuration $C^*_t(B)$ by minimizing the objective function

$$\widehat{f}(C_t) \quad = \quad \sum_{z_{i,t} \in G_t \subseteq F_t} \left\| p_{i,t}\left(C_t, R_{t-1}\right) - z_{i,t} \right\|^2 + l\left(C_t\right),$$

for each time-step $t \in \{1, \ldots, T\}$. Here, $l(C_t)$ stands for a quadratic joint limit cost term and $G_t \subseteq F_t$ denotes the set of observations $z_{i,t}$ that are associated to a marker by the Hungarian method [19]. Schubert et al. [25] and Meyer et al. [20] suggested to approximate $P\left(C^*_{1:T}(B) \mid F_{1:T}\right)$ by

$$P\left(C^*_{1:T}(B) \mid F_{1:T}\right) \quad \approx \quad \eta \prod_{t=1}^{T} \exp(-\widehat{f}(C^*_t(B))), \quad (3)$$

where η denotes a normalization factor. Accordingly, the necessity to keep the number of unlabeled markers $|F_t \setminus G_t|$ low is considered only in the association step and not in the optimization function. Therefore Eq. (3) is appropriate in the case of known and fixed bone parameters, but unsuitable for the general case. In the following section, we present a different approximation of the problem specified in Eq. (1) that takes the marker association into account.

IV. Optimization Function

Despite being more general, our improved approximation of $P\left(C_{1:T}(B) \mid F_{1:T}\right)$ relies on three assumptions. Firstly, we assume that the markers are attached to the skin, coat or cloth of the mammal to be tracked. This ensures that the skeleton structure is located somehow inside the marker cloud. We consider this assumption by adding a cylindrical error term CY. Secondly, we assume that each observed marker position corresponds to a marker attached to the mammal. Accordingly, we introduce a function UO, which computes the percentage of unlabeled observations. Thirdly, we assume that markers are attached to each modeled bone, which we take into account with a function COV (coverage). In summary, we approximate the left-hand side term of Eq. (3) by the equation

$$P\left(C^*_{1:T}(B) \mid F_{1:T}\right) \approx \eta \prod_{t=1}^{T} \exp\left(-\alpha \cdot F\left(C^*_t(B)\right)\right), \qquad (4)$$

where $\alpha = (\alpha_1, \ldots, \alpha_4)^T \in \mathbb{R}^4$ denotes a weight vector chosen manually beforehand and $F = (\widehat{f}, \mathrm{CY}, \mathrm{UO}, \mathrm{COV})^T$ denotes the new $\mathbb{R}^4$-valued performance function. The weight vector is needed to respect the different scalings and can be set once for all different mammals. In the following we give a more detailed definition of the new parts in the optimization function.

A. Cylindrical Error Function

As mentioned above, this error function ensures that the skeleton structure is located inside the marker cloud. For a given skeleton configuration C_t, we compute the likelihood that the set of observed positions F_t is attached to the body given by the skeleton configuration C_t through

$$\prod_{z \in F_t} L\left(z \mid C_t\right), \qquad (5)$$

where we compute the individual likelihoods $L\left(z \mid C_t\right)$ as in Meyer et al. [20]. Namely, assume first that you know the whole anatomy $A(C_t)$ of the mammal, i.e., the global position and shape of all muscles, given the skeleton configuration C_t, then the likelihood $L\left(z \mid C_t\right)$ decreases exponentially in the squared distance $d^2_{A(C_t)}(z)$ from the anatomy. Formally, one can formulate this as

$$L\left(z \mid C_t\right) = \eta \cdot \exp(-d^2_{A(C_t)}(z)).$$

Since the whole anatomy A of the mammal is not known, we approximate it by a concatenation of cylinders around the skeleton configuration. To be more exact, we identify C_t with a subset in $\mathbb{R}^3$, namely a union of bones B_j, i.e., $C_t = \bigcup_{j=1}^{N} B_j$ and we define $d_{C_t}(z)$ to be the distance of the marker position $z \in F_t$ to the nearest bone, denoted by $B_{\widehat{j}} = B_{\widehat{j}(z)}$ of C_t. Formally,

$$\begin{aligned} d_{C_t}(z) \quad &= \quad \min_{p \in C_t} \|z - p\| \\ &= \quad \min_{p \in B_{\widehat{j}(z)}} \|z - p\|. \end{aligned}$$

Then the likelihood is approximately given by

$$L\left(z \mid C_t\right) \quad \approx \quad \eta \cdot \exp\left(-(d_{C_t}(z) - W_{\widehat{j}(z)})^2\right), \quad (6)$$

where $W_{\hat{j}}$ denotes the width of the muscle of bone $B_{\hat{j}}$. Putting all together, we obtain the cylindrical error function

$$\mathrm{CY}(C_t) \quad = \quad \sum_{z \in F_t} \left(d_{C_t}(z) - W_{\hat{j}(z)} \right)^2 . \tag{7}$$

B. Unlabeled Observations

The next error function ensures that each observed marker is associated to a segment. In each time step t the skeleton tracking algorithm has to associate the marker observations to markers. In some cases not every marker observation can be associated to a corresponding marker with high certainty. In such cases, we experimentally found that it is better to ignore these markers rather than incorporating a wrong association. Still, it is desirable to reduce the number $N_{M,u,t}$ of unlabeled marker observations. Let $N_{M,total,t}$ be the total number of marker observations at time-step t. Then we compute the percentage of unlabeled marker observations UO through $\mathrm{UO}(C_t) = \frac{N_{M,u,t}}{N_{M,total,t}}$.

C. Coverage Function

We also use an error function that ensures that markers are attached to every bone in the model. In practice, each marker needs to be associated to a corresponding segment. Meyer et al. [20] did this once in the initialization step, while Schubert et al. [25] allowed this association to change over time so that errors in the initialization can be corrected. Thus, we define the coverage COV as the percentage of segments without an attached marker. This completes the definition of Eq. (4) and we will focus in the following section on solving the global optimization problem given in Eq. (2).

V. Bayesian Optimization

An exhaustive search over all possible skeletons or acquiring the ground truth is often unfeasible. The high-dimensional optimization problem (24 dimensions for sheep, 18 for humans) defined in Eq. (2) is non-convex as it possesses many local minima and discontinuities due to the discrete nature of the association step. Furthermore, association errors have a big influence on the tracking performance.

In this paper, we solve this global optimization problem using Bayesian optimization. Bayesian optimization has recently been successfully applied to optimize robot gait parameters [7, 8, 29], machine learning algorithms in general [4, 27, 30], and deep learning algorithms in particular [5, 11]. In the following we briefly recap it.

Bayesian optimization (see [6] for a detailed exposition) aims to find the minimum of a function $f : \mathbf{\Lambda} \rightarrow \mathbb{R}$ that is expensive to evaluate, non-convex and potentially discontinuous, based solely on (potentially noisy) function evaluations, i.e., we have to solve

$$\underset{\boldsymbol{\lambda} \in \mathbf{\Lambda}}{\operatorname{argmin}} f(\boldsymbol{\lambda}). \tag{8}$$

In order to do so, it iteratively fits and updates a probabilistic regression model $\mathcal{M}$ of f based on previous point evaluations of f and any prior information available, and uses this model to select the next next point $\boldsymbol{\lambda}$ to evaluate. It does so by trading off exploration (evaluating f in a region of the parameter space in which $\mathcal{M}$ is uncertain) versus exploitation (evaluating f at points $\mathcal{M}$ predicts to yield low function values). This exploration/exploitation trade-off is formalized by means of a so-called *acquisition function*. Here, we explain the most popular acquisition function of *expected improvement* (EI) [24] over the best function value f_{min} observed so far. In our experiments we also tested the *upper confident bound* (UCB) [28] as an acquisition function and obtained similar results.

Let $I_{f_{min}}(\boldsymbol{\lambda}) = \max(0, f_{min} - f(\boldsymbol{\lambda}))$ be the *positive improvement* over f_{min} at a parameter setting $\boldsymbol{\lambda}$; the expected improvement is then simply the expectation over $I_{f_{min}}(\boldsymbol{\lambda})$, taken with respect to the predictive distribution of $\mathcal{M}$, i.e.,

$$\mathbb{E}_{\mathcal{M}}[I_{f_{min}}(\boldsymbol{\lambda})] = \int_{-\infty}^{f_{min}} \max\{f_{min} - f, 0\} \cdot p_{\mathcal{M}}(f \mid \boldsymbol{\lambda}) \, df. \tag{9}$$

In the case of Gaussian predictive distributions with predictive mean $\mu_{\boldsymbol{\lambda}}$ and standard deviation $\sigma_{\boldsymbol{\lambda}}$, this integral can be solved in closed form as

$$\mathbb{E}_{\mathcal{M}}[I_{f_{min}}(\boldsymbol{\lambda})] = \sigma_{\boldsymbol{\lambda}} \cdot [u \cdot \Phi(u) + \varphi(u)], \tag{10}$$

where $u := \frac{f_{min} - \mu_{\boldsymbol{\lambda}}}{\sigma_{\boldsymbol{\lambda}}}$ and φ and Φ denote the probability density function and cumulative distribution function of a standard normal distribution, respectively.

In order to apply Bayesian optimization for the maximization problem Eq. (2), we equivalently minimize the negative logarithmic of Eq. (4) over the bone parameter vector $B =: \mathbf{\Lambda}$. Thus we need to minimize

$$f(B) = \sum_{t=1}^{T} \alpha \cdot F\left(C_t^*(B)\right). \tag{11}$$

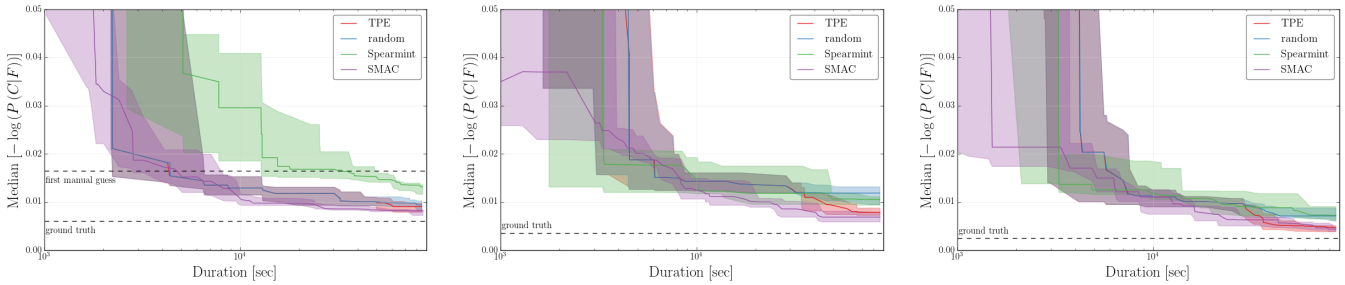
When we have multiple recordings of an individual we may also want to optimize their bone parameter vector jointly across all of these recordings. Given K recordings $F_{1:T}^{(1)}, \ldots, F_{1:T}^{(K)}$, we simply minimize the average loss g across several functions $f_1, \ldots, f_K$, i.e.,

$$g(B) = \frac{1}{K} \sum_{k=1}^{K} f_k(B), \tag{12}$$

with f_k using data $F_{1:T}^{(k)}$.

Some Bayesian optimization methods can exploit this additive substructure by evaluating the performance on one recording at a time; when performance is poor, this allows to stop the evaluation and save time that can be used to evaluate more parameter settings in the same computational budget. Other differences in Bayesian optimization methods are mostly related to the model classes they use. In order to demonstrate that several approaches can be used in our context, we evaluated the following three popular instantiations (using the unified interfaces provided by the library HPOlib [13]):

Sequential Model-based Algorithm Configuration (SMAC) [17] uses a random forest to model $p_{\mathcal{M}}(f \mid \boldsymbol{\lambda})$ and is the only SMBO method that implements a mechanism



(a) Sheep, 9 recordings (b) male human, 8 recordings (c) female human, 9 recordings

Fig. 2: Tracking error of the best found skeleton over time. We show median and upper/lower quantiles across 10 runs of each optimizer and random search.

to exploit the additive substructure when optimizing across recordings. While other optimizers need to evaluate each parameter setting on all recordings, SMAC evaluates only a single recording at a time and rejects bad configurations as early as possible.

Spearmint [27] implements standard Bayesian optimization and uses a Gaussian Process (GP) to model $p_{\mathcal{M}}(f \mid \boldsymbol{\lambda})$, using slice sampling over the GP's hyperparameters.

Tree Parzen Estimator (TPE) [4] in contrast models $p(f < f*)$, $p(f \mid \boldsymbol{\lambda})$, and $p(\boldsymbol{\lambda} \mid f \geq f*)$, with a tree-structured Parzen density estimator where $f*$ is a fixed quantile of the losses observed so far. With these distributions a term proportional to Eq. (9) can be computed in closed form.

VI. Experimental Results

We collected our sheep recordings in a barn of width 3 m and length 8 m. We restricted the number of used markers to 40, which led to a sparse coverage so that on some modeled bones only one or two markers are attached to. Due to the restricted area many markers are occluded if the sheep stands in a corner of the barn. The skeleton tracking methods defined by Ringer and Lasenby [23], Kirk et al. [18], and de Aguiar et al. [12] require connected sequences of marker frames over longer periods of time or they need time-consuming preprocessing steps. Thus they are not reasonably applicable for the datasets we collected.

We use ten Raptor-E cameras at 100 Hz to record datasets for the experiments. Our datasets consist of nine recordings of the sheep, nine recordings of a female and eight recordings of a male. Each recorded dataset consists of at least one minute of motion data. For the sheep we considered a 24-dimensional optimization problem, and for the human dataset we optimized over 18 dimensions. We ran all experiments on Intel Xeon E5-2650 v2 CPUs, where ten simultaneous optimization runs shared 16 cores and 64 GiB RAM. Evaluating our performance function (see Eq. (11)) on one bone parameter configuration and on one recording took between 150 and 250 sec.

In order to score our obtained skeleton structure in an offline analysis phase, we need to know the ground truth skeleton structure C_{GT}, which is unavailable in practice. For the human datasets we approximated the ground truth skeleton by the skeleton structure given by Contini [9]. For the sheep recordings we first took the skeleton from a sketch, which we denote in the following as first manual guess. Then we measured the bone lengths of the narcotized sheep and modified them iteratively to fit into the marker cloud and to obtain good tracking results. In the following, we call the skeleton we obtained by this time-consuming iterative procedure "ground truth skeleton".

The following experiments show the robustness of our automated method.

A. Global Optimization

To study whether Bayesian optimization can automatically find a good skeleton structure, we first need to define the search space. For each free parameter, we choose a range of $\frac{1}{2}$ to 2 times the parameter's value in the ground truth skeleton. We assume that a human can guess such a rough range for the parameter vector of the mammal to be tracked, without manual measurements or other time-consuming analysis techniques.

Tracking performance. For each of our three datasets, we ran each of our three Bayesian optimization techniques (SMAC, TPE, Spearmint) for 24h, as well as a baseline optimizer based on random search [3]. In order to quantify the uncertainty in our results we performed ten independent runs of each of these methods. For the optimizer SMAC we evaluated next to EI (Eq. (10)) also the acquisition function UCB [28] like Calandra et al. [8]; in our case, this yielded qualitatively similar results. We omit the results in the figure to avoid clutter.

TABLE I: Means across 10 runs of each optimizer. For each row, bold face indicates the best mean loss, and underlined values are not statistically significantly different from the best according to an unpaired t-test (with p=0.05). Additionally, we report performance of the ground truth (gr.truth) and for the sheep first manual guess (manual).

	#record.	**SMAC**	**Spearmint**	**TPE**	gr.truth	manual
Male	8	**0.0070**	0.0120	0.0080	0.0036	-
Female	9	**0.0045**	0.0093	0.0047	0.0025	-
Sheep	9	**0.0077**	0.0123	0.0090	0.0061	0.0165

Fig. 2 shows the tracking error of the best parameter configuration each method found over time, compared to tracking performance with the ground truth skeleton. We further provide quantitative results for these experiments in Table I. All optimizers found parameter vectors that substantially improved tracking performance over the initial guess and almost reached the performance of the ground truth skeleton with a completely automated process that had no knowledge of the ground truth.

Comparing the performance of the individual optimizers, we note that the tree-based Bayesian optimization methods SMAC and TPE yielded the best performance, with SMAC having a slight advantage because it can evaluate parameter settings based on the performance they yield on individual recordings (while the other methods always have to evaluate all recordings for a new parameter configuration). The Gaussian-Process-based method Spearmint is known to work well for low-dimensional optimization problems but to have problems in higher dimensions [13]; indeed, here it performed similar to the baseline (random search) for the 18-dimensional human datasets and performed even worse than the baseline for the 24-dimensional sheep dataset. Based on these results we will only consider the optimizer SMAC for all further experiments.

Distance to ground truth skeleton. We confirm this finding by comparing the best found skeleton by the optimizer SMAC with the ground truth skeleton. For this purpose, we calculate the joint position difference between the obtained optimal skeleton and the ground truth. In Table II we show mean and standard deviation of the differences and the maximal value of the joint position differences. We compare the skeletons obtained by optimizing with respect to our proposed performance measure in Eq. (4) and with respect to the performance measure introduced by Meyer et al. [20] (Eq. (3)). The result shows that using our objective function, which allows unlabeled observations and considers the coverage and cylindric error, yields skeletons closer to the ground truth.

The main reason for the larger distances for the sheep in comparison to the humans are sub-optimal marker placements, which do not allow joint position inference. It was not possible to attach markers to all bones of the sheep and additionally they shifted towards the joints.

TABLE II: Joint position difference $[cm]$

	Meyer et al. [20] Eq. (3)			OUR Eq. (4)		
	mean	**std**	**max**	**mean**	**std**	**max**
Male	2.09	0.92	4.17	1.18	0.65	2.27
Female	2.68	1.78	5.75	1.11	0.67	2.26
Sheep	9.81	4.45	18.86	8.15	4.06	11.88

Smoothness. Although smooth motion trajectories are desirable in motion capture, we did not incorporate smoothness in our performance function to avoid biasing skeleton movements towards motions with zero movement. Our results show that the optimized skeleton structure nevertheless leads to smoother motion trajectories than a manually-determined one. In Fig. 3 we plot the velocity of the root segment of the sheep for the approximate ground truth skeleton, the optimized one and a first manual guess for the skeleton structure. We obtained similar results with all other joints and motion sequences in our datasets.

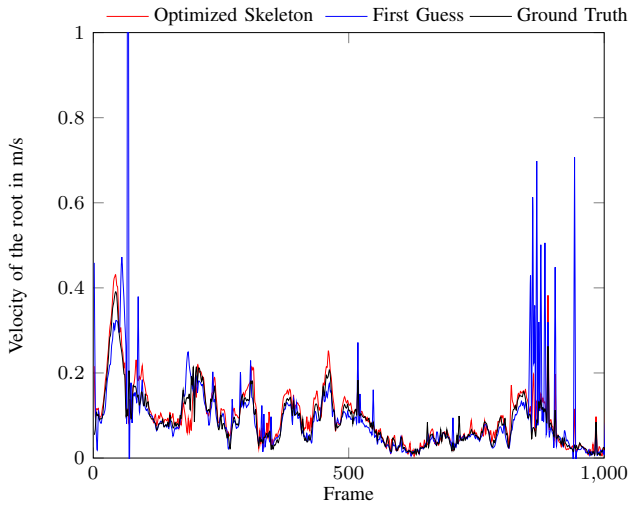


Fig. 3: Sheep: Smoothness of the velocity of the root segment obtained with our method in comparison to the ground truth skeleton and a first guess. Absolute differences to ground truth over all datasets: 0.0252± 0.0641 m/s (First Guess), 0.0177± 0.0325 m/s (Optimized Skeleton)

B. Start Values

As mentioned above, many markers are occluded if the sheep stands in a corner of the barn and thus the first frame is not the optimal one for initialization of the skeleton structure. Since manually defining the right start frame is a time-consuming task, we also incorporate the start value in our search space and optimize over each dataset individually. We choose the start value in the range of 1000 frames before, which we call first frame in Fig. 4, and 500 frames after the manual chosen one. We note, that this does not change the range of frames across which we calculate the tracking performance. Fig. 4 shows the velocity profile for one dataset and the optimized skeleton, the ground truth skeleton started at the first frame and the ground truth started at the manually chosen one. We obtained similar results for all our datasets.

VII. Conclusion and Future Work

In this paper, we presented an approach to obtain a skeleton of a mammal out of moving sequences of unlabeled marker positions in an optical motion capture system. Our approach is based on a probabilistic skeleton tracking approach and employs a novel approximation for the case of online skeleton tracking, which takes unlabeled observations, coverage and the anatomy into account. We showed that Bayesian optimization with respect to this approximation robustly yields good skeletons based on a rough initial guess.

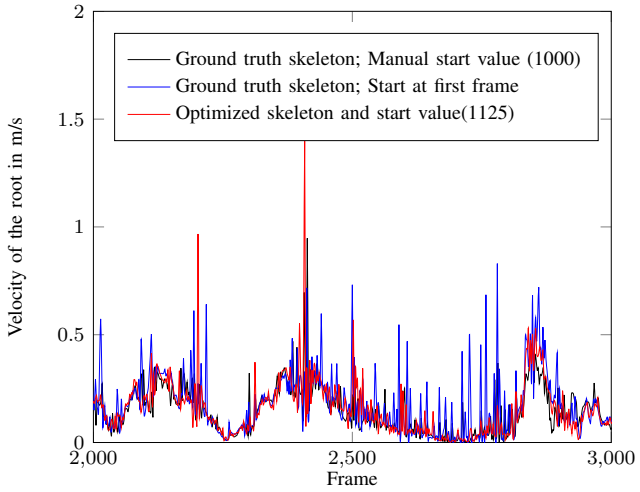


Fig. 4: Sheep: Smoothness of the velocity of the root dependent on the start value. Absolute differences to ground truth over all datasets: 0.0635± 0.0939 m/s (First Guess), 0.0343± 0.0571 m/s (Optimized Skeleton)

Our method also automatically finds a suitable start frame for initializing the tracking. In the future, we plan to record more mammals to evaluate how well this approach generalizes to additional mammals and partial views.

References

[1] O. K.-C. Au, C.-L. Tai, H.-K. Chu, D. Cohen-Or, and T.-Y. Lee. Skeleton extraction by mesh contraction. *ACM Trans. on Graph*, 27, 2008.

[2] A. Baak, M. Müller, G. Bharaj, H. P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *ICCV'11*, 2011.

[3] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *JMLR*, 13, 2012.

[4] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Proc. of NIPS'11*, 2011.

[5] J. Bergstra, D. Yamins, and D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proc. of ICML'13*, 2013.

[6] E. Brochu, V. Cora, and N. de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *CoRR*, 1012.2599, 2010.

[7] R. Calandra, N. Gopalan, A. Seyfarth, J. Peters, and M. Deisenroth. Bayesian gait optimization for bipedal locomotion. In *Proc. of LION-8*, 2014.

[8] R. Calandra, A. Seyfarth, J. Peters, and M. Deisenroth. An experimental comparison of Bayesian optimization for bipedal locomotion. In *Proc. of ICRA'14*, 2014.

[9] R. Contini. Body segment parameters II. *Artificial Limbs*, 16 (1), 1972.

[10] Motion Analysis Corp. *Cortex*, 2013. URL http://www.motionanalysis.com/html/industrial/cortex.html. Version 4.0.0.1387.

[11] G. Dahl, T. Sainath, and G. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *Proc. of ICASSP'13*, 2013.

[12] E. de Aguiar, C. Theobalt, and H.-P. Seidel. Automatic learning of articulated skeletons from 3D marker trajectories. In *Second Int. Symposium on Advances in Visual Computing (ISVC), Part I*, 2006.

[13] K. Eggensperger, M. Feurer, F. Hutter, J. Bergstra, J. Snoek, H. Hoos, and K. Leyton-Brown. Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. In *NIPS Workshop on Bayesian Optimization in Theory and Practice*, 2013.

[14] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[15] B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester. Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease. *Gait & Posture*, 39, 2014.

[16] W. Huang, W. Shihao, D. Cohen-Or, M. Gong, H. Zhang, G. Li, and B. Chen. L1-medial skeleton of point cloud. *ACM Trans. on Graph*, 32, 2013.

[17] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proc. of LION-5*, 2011.

[18] A. G. Kirk, J. F. O'Brien, and D. A. Forsyth. Skeletal parameter estimation from optical motion capture data. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2005.

[19] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2), 1955.

[20] J. Meyer, M. Kuderer, J. Müller, and W. Burgard. Online marker labeling for fully automatic skeleton tracking in optical motion capture. In *Proc. of ICRA'14*, 2014.

[21] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2–3), 2006.

[22] S. Obdrzalek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel. Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population. In *IEEE Conf. on Engineering in Medicine and Biological Society (EMBC'12)*, 2012.

[23] M. Ringer and K. Lasenby. A procedure for automatically estimating model parameters in optical motion capture. In *Proc. of the British Machine Vision Conference*, 2002.

[24] M. Schonlau, W. J. Welch, and D. R. Jones. Global versus local search in constrained optimization of computer models. In *New Developments and Applications in Experimental Design*, volume 34. 1998.

[25] T. Schubert, A. Gkogkidis, T. Ball, and W. Burgard. Automatic initialization for skeleton tracking in optical motion capture. In *Proc. of ICRA'15*, 2015.

[26] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab. Human skeleton tracking from depth data using geodesic distances and optical flow. *Image and Vision Computing*, 30 (3), 2012.

[27] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Proc. of NIPS'12*, 2012.

[28] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. of ICML'10*, 2010.

[29] M. Tesch, J. Schneider, and H. Choset. Using response surfaces and expected improvement to optimize snake robot gait parameters. In *Proc. of IROS'11*, 2011.

[30] C. Thornton, F. Hutter, H. Hoos, and K. Leyton-Brown. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. In *Proc. of KDD'13*, 2013.

[31] J. Yan and M. Pollefeys. A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30, 2008.