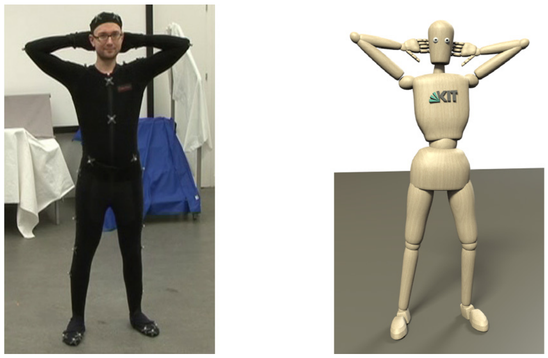
Real-Time Whole-Body Human Motion Tracking Based on Unlabeled Markers

Jannik Steinbring[1], Christian Mandery[2], Florian Pfaff[1], Florian Faion[1], Tamim Asfour[2], and Uwe D. Hanebeck[1]

***Abstract*— In this paper, we present a novel *online approach* for tracking whole-body human motion based on *unlabeled measurements* of markers attached to the body. For that purpose, we employ a given kinematic model of the human body including the locations of the attached markers. Based on the model, we apply a combination of constrained sample-based Kalman filtering and multi-target tracking techniques: 1) joint constraints imposed by the human body are satisfied by introducing a parameter transformation based on periodic functions, 2) a global nearest neighbor (GNN) algorithm computes the most likely one-to-one association between markers and measurements, and 3) multiple hypotheses tracking (MHT) allows for a robust initialization that only requires an upright standing user. Evaluations clearly demonstrate that the proposed tracking provides highly accurate pose estimates in real-time, even for fast and complex motions. In addition, it provides robustness to partial occlusion of markers and also handles unavoidable clutter measurements.**

I. Introduction

Knowledge about whole-body human motion is a key ingredient for a large number of research areas, including the field of computer graphics and animation, robotic applications, e.g., imitation learning, biomechanical analysis, e.g., gait analysis for rehabilitation, and human intention recognition. The most intuitive and comprehensive way to acquire such human motion is to track the whole-body movements performed by a subject. In addition, for certain applications, it is desirable to acquire the motion in real time, e.g., to directly inspect the reconstructed motion or reproduce it on a humanoid robot. An established and widely used way of capturing human motion is to use commercial marker-based motion capture systems, such as Vicon systems, which can provide discrete-time position measurements of non-unique/unlabeled markers attached to the human body (see Fig. 1a). In order to gain knowledge about the motion from these noisy marker trajectories, they can be used to determine the time-varying parameters of a kinematic model, i.e., joint angle values, and root position and orientation, that describe the human pose.



(a) Optical motion capture. (b) Estimated human pose.

Fig. 1: The proposed whole-body motion tracking. Unlabeled markers attached to the human body are measured using optical motion capture and used to estimate the human pose.

Due to the nonlinear relationship between the marker measurements and the model parameters to be determined, the considered tracking is equivalent to estimating the state of a discrete-time stochastic nonlinear dynamic system, where the system state is the human pose (see Fig. 1b). Popular recursive state estimators are (nonlinear) Kalman filters [1], [2] or particle filters [3]. The advantage of such estimators is that they maintain a probability distribution of the state estimate and use this distribution to optimally fuse the current state estimate with newly available noisy measurements to obtain an updated distribution.

A. Contribution

In [4], we proposed a real-time whole-body motion tracking using *labeled* marker measurements based on recursive nonlinear state estimation. In this paper, we extend our approach to the much more complicated case of *unlabeled* marker measurements. For that purpose, we assume a known kinematic model of the human body, where joint angles and root pose are the only time-varying parameters. This also includes joint limits derived from biomechanical analysis and the locations of the markers attached to the body.

Unfortunately, due to the many degrees of freedom (DoF) required for a detailed kinematic model, particle filters are not suitable for real-time tracking as they would need a huge amount of particles to get meaningful state estimates. Hence, we use the smart sampling Kalman filter (S^{2}KF) [5] to allow for estimating all parameters of the kinematic model in real time. Moreover, the imposed joint limits turn the tracking task into a constrained estimation problem. In order to satisfy these limits, we transform the constrained estimation problem into an unconstrained problem with the aid of periodic functions. This is necessary, as Kalman filters, including the S^{2}KF, can only estimate unconstrained quantities.

The main challenge, however, arises from the unlabeled marker measurements, as they form an instance of multi-target tracking with unknown associations. More precisely, the

[1] Authors are with the Intelligent Sensor-Actuator-Systems Laboratory (ISAS), Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT), Germany. E-mail: jannik.steinbring@kit.edu, florian.pfaff@kit.edu, florian.faion@kit.edu, uwe.hanebeck@ieee.org

[2] Authors are with the High Performance Humanoid Technologies Lab (H^{2}T), Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT), Germany. E-mail: mandery@kit.edu, asfour@kit.edu

The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement no. 611909 (KoroiBot) and the European Union H2020 Programme under grant agreement no. 643666 (I-SUPPORT).

markers attached to the body are *targets* and for the measured locations, it is not known from which marker they originate. Our solution to this is twofold. On the one hand, a global nearest neighbor (GNN) approach [6, Sec 6.4], [7, Sec. 10.3.1] is used to compute optimal associations between observed marker locations and predicted marker locations. On the other hand, a customized version of multiple hypotheses tracking (MHT) [6] is used to maintain multiple *pose hypotheses* over time. This is required to cope with convergence to wrong poses, which is unavoidable due to the unknown associations and especially challenging when markers are observed for the first time.

B. *Related Work*

The authors of [8] proposed a method to track movements of human hands with unlabeled markers that also uses a GNN approach in the form of the Hungarian algorithm. In addition, they deal with marker occlusions by performing interpolation procedures. As opposed to this, we perform motion tracking for the entire body. Moreover, we consequently model uncertainties, and thanks to the use of recursive state estimation techniques, our tracking can naturally deal with marker occlusions without any special interpolation procedure.

In [9], whole-body motion tracking based on unlabeled markers is considered as well. Here, also the Hungarian algorithm is used for an one-to-one assignment of markers and measurements, and a known kinematic model of the human body is assumed that, however, does *not* include the locations of the attached markers. Instead, the locations of the markers and the corresponding body segments are estimated during the initialization phase of the tracking. Compared to our approach this has some drawbacks. First, usually anatomically inspired marker placements are used and markers are not placed at arbitrary locations. Such information is not utilized in this approach, and hence, lost. Second, each subject is required to perform a special initialization pose, i.e., a T-pose, for the actual tracking.

Furthermore, the work in [9] is extended in [10] with a faster initialization procedure using the k-means algorithm. However, the implementation is tailored to a sheep, e.g., assuming a number of four or five marker clusters and providing a heuristics for detecting the head markers, and has not been evaluated with a human subject. Furthermore, the authors provide an MHT approach for an automatic initialization based on a large set of already captured motions. Unfortunately, this initialization can take several minutes, making their approach intractable for real-time tracking. In contrast to this, our tracking does not need such a special initialization phase, and can operate in real time a few seconds after the tracking started. We only assume that the human initially stands upright. For example, in our approach the human can simply walk into the area observed by the marker-based motion capture system to start the tracking automatically. This makes our proposed whole-body motion tracking very user-friendly.

The problem of missing marker positions is addressed in [11]. Their approach is to predict missing marker positions using previously known marker positions and to get information based on rigid body assumptions. Instead, our proposed recursive state estimation approach implicitly takes information of previous frames into account to be able to handle missing markers more easily. Moreover, in [11] no fixed kinematic model is used, since they do online joint localization in the marker point cloud, which results in a time-varying kinematic model.

II. The Master Motor Map Framework

The Master Motor Map (MMM) framework [12], [13] provides an open-source framework for capturing, representing, and analyzing human motion and reproducing it on humanoid robots. At its core, it provides the MMM reference model, a whole-body model for the human body based on well-established biomechanics literature that can be scaled to the measured body height of a human subject. This reference model can represent human motion using 6 DoF for the root pose, 52 DoF for torso, extremities, head, and eyes, and 2 × 23 DoF for the fingers of both hands. It is explained in more detail in [12] and exemplified in Fig. 1b. A reference marker set that can be used for whole-body motion capture (specifications are given in [14]) is also part of the MMM reference model.

The MMM framework already provides procedures for the reconstruction of human motion, which formulate the problem as a frame-wise optimization problem [12], [15]. However, these algorithms can only work with *labeled* marker measurements and cannot handle missing marker measurements. Hence, time-consuming and error-prone post-processing of the recorded marker trajectories is necessary to make these approaches work. In the next section, we introduce our new approach that does not have these limitations, i.e., it can be directly applied to the unlabeled marker measurements provided by a marker-based motion capture system.

III. Whole-Body Human Motion Tracking with Unlabeled Markers

In this section, we describe a novel approach for tracking whole-body motions with unlabeled markers. We start with some preconditions. At each discrete time step k, the kinematic model is characterized by J joint angles[1]

$$\underline{\theta}_k = [\theta_k^{(1)}, \ldots, \theta_k^{(J)}]^\top \qquad (1)$$

as well as by the root pose with its position $\underline{r}_k = [r_k^x, r_k^y, r_k^z]^\top$ in Cartesian coordinates and its orientation $\underline{o}_k = [o_k^r, o_k^p, o_k^y]^\top$ in roll, pitch, and yaw angles. In addition, like human joints, all joint angles are limited to an individual range

$$l_j \leq \theta_k^{(j)} \leq u_j \ \ , \ \ \forall j \in \{1, \ldots, J\} \ \ , \qquad (2)$$

where l_j denotes the lower bound and u_j the upper bound. Based on a known kinematic model of the human body including the locations of N markers attached to it, e.g.,

[1] Vectors are underlined and matrices are printed bold face.

the MMM reference model, for a given root pose and joint angles, we can compute the position $\underline{p}_k^{(n)}$ of the n-th marker in Cartesian world coordinates using the forward kinematics

$$\underline{p}_k^{(n)} = \underline{h}^{(n)}(\underline{r}_k, \underline{o}_k, \underline{\theta}_k) \ , \ \forall n \in \{1, \ldots, N\} \ . \qquad (3)$$

Furthermore, a marker-based motion capture system provides us with a set $\mathcal{M}_k = \{\underline{\tilde{m}}_k^{(1)}, \ldots, \underline{\tilde{m}}_k^{(M_k)}\}$ of M_k noisy and unlabeled marker measurements in Cartesian coordinates. Note that, due to possible occlusions and/or clutter, M_k can be smaller or larger than N.

A. Satisfying the Joint Angle Bound Constraints

Our goal is to infer, i.e., estimate, the kinematic model parameters $\underline{\theta}_k$, $\underline{r}_k$, and $\underline{o}_k$ from the received marker positions $\mathcal{M}_k$ using a sample-based Kalman filter. This estimation task, however, poses an additional challenge, as Kalman filters by design can only estimate unconstrained quantities. That is, estimating $\underline{\theta}_k$ directly with a Kalman filter may violate the constraints (2). Recall that the system state estimates of a Kalman filter are represented by Gaussian distributions, i.e., by a mean vector and a covariance matrix. In order to take the bound constraints properly into account, it is necessary that (i) the mean vector must always lie inside a bounded region of the state space, and (ii) the covariance matrix has to reflect that the state space is bounded by being smaller compared to an unconstrained state space. In literature, there exist various approaches to incorporate constraints into Kalman filters.

- *Perfect measurements* [1] are designed for equality constraints and are not suitable for inequality constraints. Hence, they cannot be applied to the considered bound constraint problem.
- *Projection techniques* [1] correct the posterior state mean after a Kalman filter prediction/update step. Unfortunately, they cannot correct the posterior state covariance matrix as well.
- *PDF truncation* [1] is an elegant way to respect linear inequality constraints and corrects both posterior state mean and covariance matrix. However, it is computationally expensive for high-dimensional states as it requires several Gram–Schmidt orthogonalizations and eigendecompositions of the state covariance matrix, which are not guaranteed to converge, and hence, make this approach unreliable.
- The *sampling-based approach* proposed in [16] can be seen as a numerical approximation of the PDF truncation approach. The problem here is that due to the large state space situations frequently occur in which too many samples lie outside of the constrained region and no constrained estimate can be computed. This is analogous to the well-known sample degeneracy problem of particle filters.

As we seek a real-time capable and accurate human motion tracking method, we choose another way to satisfy (2) for all joint angles. We perform a parameter transformation using a periodic function $g : \mathbb{R} \to [-1, 1]$. We introduce a new joint parameter $\Theta_k^{(j)}$ for each joint angle $\theta_k^{(j)}$ according to the mapping

$$\theta_k^{(j)} = g_j(\Theta_k^{(j)}) = \frac{u_j - l_j}{2} \sin(\Theta_k^{(j)}) + \frac{l_j + u_j}{2} \ .$$

As a result, $\Theta_k^{(j)}$ can take any value in $\mathbb{R}$ while (2) is always satisfied. It should be noted that this periodic approach, however, is sensitive to large uncertainties in the parameters $\Theta_k^{(j)}$, that is, their uncertainties should not be larger than the period of the periodic function to get meaningful estimation results. Alternatively, sigmoid functions like the hyperbolic tangent could also be used for such a transformation. However, experiments showed that then the filter exhibits problems to properly update a joint angle estimate in situations where it is close to a bound constraint, as the gradient of a sigmoid function becomes very small for large parameters.

Analogously to the vector $\underline{\theta}_k$ (1), we define the joint parameter vector $\underline{\Theta}_k = [\Theta_k^{(1)}, \ldots, \Theta_k^{(J)}]^\top$. We also introduce the vector-valued function

$$\underline{\theta}_k = \underline{g}(\underline{\Theta}_k) = \left[g_1(\Theta_k^{(1)}), \ldots, g_J(\Theta_k^{(J)})\right]^\top$$

that transforms all joint parameters back to their corresponding joint angles.

At this point, we can define the system state vector

$$\underline{x}_k = [\underline{r}_k^\top, \underline{o}_k^\top, \underline{\Theta}_k^\top]^\top \in \mathbb{R}^D$$

with $D = 6 + J$ that fully describes the constrained whole-body pose at time step k. This state vector can now be recursively estimated with a sample-based Kalman filter consisting of the usual alternating state prediction and measurement update.

B. State Prediction

For the state prediction, we have to model possible changes in the human's pose from one time step to the next one. Fortunately, marker-based motion capture systems work with high frame rates (100 Hz in our case), and hence, the pose will only change slightly between time steps. Hence, it is sufficient to employ the simple identity system model

$$\underline{x}_k = \underline{x}_{k-1} + \underline{w}_k \ , \qquad (4)$$

where $\underline{w}_k$ is zero-mean white Gaussian noise with covariance matrix $\mathbf{Q}_k$. Given the state mean $\underline{\hat{x}}_{k-1}^e$ and state covariance $\mathbf{C}_{k-1}^e$ from the last time step $k-1$, the predicted state mean $\underline{\hat{x}}_k^p$ and the predicted state covariance $\mathbf{C}_k^p$ can be simply computed in closed-form.

C. From State to Marker Positions

In order to update the predicted state estimate, we first need mappings from $\underline{x}_k$ to each individual marker position. Those mappings consist of two parts. On the one hand, given a specific system state, for each marker the forward kinematics of the respective kinematic chain has to be computed using (3). As a result, it is known where to *expect* all markers for the human pose described by the respective system state. On the other hand, the measured marker positions are subject to noise. Hence, uncertainty has to be taken into account in order to obtain good estimation results, especially in case of high noise. Both together leads to the desired mappings

$$\left.\begin{aligned}\underline{m}_k^{(n)} &= \underline{h}^{(n)}(\underline{x}_k) + \underline{v}_k^{(n)} \\ &= \underline{h}^{(n)}(\underline{r}_k, \underline{o}_k, \underline{g}(\underline{\Theta}_k)) + \underline{v}_k^{(n)}\end{aligned}\right. , \forall n \in \{1, \dots, N\}, \quad (5)$$

where $\underline{v}_k^{(n)}$ is additive zero-mean white Gaussian noise with covariance matrix $\mathbf{R}_k^{(n)}$. The choice of $\mathbf{R}_k^{(n)}$ depends on the utilized tracking system. Moreover, it is assumed that the noise vectors $\underline{v}_k^{(i)}$ and $\underline{v}_k^{(j)}$ with $i \neq j$ are mutually independent, and that each noise vector $\underline{v}_k^{(n)}$ is also independent of the system state $\underline{x}_k$. The mappings (5) are used in Section III-E to construct the measurement equation that is required for the measurement update.

D. Marker–Measurement Association and Outlier Detection

Now, we have to tackle the central problem of unknown marker–measurement associations and the detection of potential measurement outliers. That is, given the predicted state estimate, i.e., $\underline{\hat{x}}_k^p$ and $\mathbf{C}_k^p$, for each received measurement, we have to decide whether it is an outlier in order to discard it and, if not, determine from which marker it originates. This task boils down to a multi-target tracking problem, where all targets move in a collaborative manner due to the underlying kinematic model. Removing measurement outliers and obtaining optimal marker–measurement associations consists of several steps.

First, for each marker $1 \leq n \leq N$ we compute the predicted position mean

$$\underline{\hat{m}}_k^{(n)} = \frac{1}{S} \sum_{s=1}^{S} \underline{h}^{(n)}(\underline{x}_k^{(s)})$$

and predicted position covariance matrix

$$\begin{aligned}(\mathbf{C}_k^m)^{(n)} = \frac{1}{S} \sum_{s=1}^{S} & (\underline{h}^{(n)}(\underline{x}_k^{(s)}) - \underline{\hat{m}}_k^{(n)}) \cdot \\ & (\underline{h}^{(n)}(\underline{x}_k^{(s)}) - \underline{\hat{m}}_k^{(n)})^\top + \mathbf{R}_k^{(n)} \quad ,\end{aligned}$$

where the equally weighted samples $\underline{x}_k^{(s)}$ approximate the prior Gaussian state estimate $\mathcal{N}(\underline{x}_k; \underline{\hat{x}}_k^p, \mathbf{C}_k^p)$ with the aid of the Gaussian sampling technique from the smart sampling Kalman filter (S^{2}KF) [5].

Second, based on the predicted marker means, measurement outliers are removed using ellipsoidal gates [6, Sec. 6.3.2], i.e., a measurement $\underline{\tilde{m}}_k^{(i)}$ is discarded if $\|\underline{\tilde{m}}_k^{(i)} - \underline{\hat{m}}_k^{(n)}\|_2 > \varepsilon_o, \forall\, 1 \leq n \leq N$. The remaining M'_k measurements are given by the set $\mathcal{M}'_k$. As the markers attached to the body can slightly move during locomotion, the common gating based on the Mahalanobis distance leads to problems, and thus, we choose to use the Euclidean distance instead.

Third, with the remaining measurements $\mathcal{M}'_k$, we determine the most probable one-to-one assignment between predicted marker positions and measurements using a GNN approach. This has the advantage that one marker will only be associated to exactly one measurement, and thus, reflects the fact that each marker *can* only generate one measurement per time step. In order to incorporate the uncertainty of the state estimate and the measurement noise into the association procedure, we compute the Mahalanobis distances between all predicted marker positions and measurements according to

$$d^{(n,i)} = (\underline{\hat{m}}_k^{(n)} - \underline{\tilde{m}}_k^{(i)})^\top \left((\mathbf{C}_k^m)^{(n)}\right)^{-1} (\underline{\hat{m}}_k^{(n)} - \underline{\tilde{m}}_k^{(i)}) \ ,$$

with $1 \leq n \leq N$ and $1 \leq i \leq M'_k$. These $d^{(n,i)}$ build the cost matrix $\mathbf{D}_k \in \mathbb{R}^{N,M'_k}$ to be minimized by the association algorithm. It is useful to understand that the association that maximizes the product of the probability densities also minimizes the sum of Mahalanobis distances [17, Sec 11.3]. Finding the association that minimizes the sum of the costs is a classical linear assignment problem (LAP), which can be solved, e.g., by the Hungarian algorithm [18]. Modern variants of the Hungarian algorithm feature a runtime complexity of $\mathcal{O}(n^3)$, and Jonker and Volgenant [19] proposed a very fast solver called LAPJV, which we utilize in our implementation.

Due to potential occlusions of markers and erroneous measurements not stemming from markers (clutter), the cost matrix $\mathbf{D}_k$ is not necessarily square, which is the expected input format for many LAP solvers such as LAPJV. Hence, if $M'_k \neq N$ we have to extend the cost matrix $\mathbf{D}_k$ to a square one. If $M'_k < N$, we introduce $N - M'_k$ "fake measurements", or if $M'_k > N$, we introduce $M'_k - N$ "fake markers". These get a cost that is larger than any distance between the actual measurements and predicted marker positions. This prevents the "fake measurements/markers" to compete with the non-fake entries, and thus, ensures that declaring a measurement as clutter or a marker as occluded is only done as the last resort.

Note that filling up the cost matrix only until it is square poses the risk of false assignments in case of simultaneous clutter and occlusions. While there are more sophisticated ways to account for this problem [6, Sec. 6], they are hard to parametrize for our scenario. Moreover, due to the preceding gating step, errors induced by simultaneous clutter and occlusions are reduced to a minimum.

Based on the constructed cost matrix, the LAPJV algorithm computes the optimal marker–measurement associations. From these associations, we only use the $A = \min(M'_k, N)$ associations with smallest costs, as we have only N markers that can be associated to a measurement, i.e., the "fake measurements/markers" are ignored. The indices of the selected measurements are $\{s_1, \dots, s_A\}$, with $1 \leq s_i \leq M'_k$ and $s_i \neq s_j$, whereas the indices of the associated markers are $\{a_1, \dots, a_A\}$, with $1 \leq a_i \leq N$ and $a_i \neq a_j$.

Please note that there are multi-target tracking approaches in literature [6], [7] that are more sophisticated or significantly faster. Greedy approaches such as the (local) nearest neighbor (LNN) [7, Sec. 10.3] can return an association in $\mathcal{O}(n^2)$ but its performance quickly deteriorates in regions where markers are densely clustered. Better alternatives for suboptimal approaches are Auction algorithms [20], which provide an upper bound for their suboptimality in the worst case. However, since the majority of computation is used for the nonlinear filtering, we deem LAPJV to be fast enough and do not need to sacrifice assignment quality for higher speed.

E. Measurement Update

Next, we need a *single* measurement vector $\underline{\tilde{m}}_k$ constructed out of the associated measurements and a measurement equation that models the relationship between the system state $\underline{x}_k$ and this constructed measurement vector. The measurement vector is constructed by stacking the selected A marker measurements according to

$$\underline{\tilde{m}}_k = [(\underline{\tilde{m}}_k^{(s_1)})^\top, \dots, (\underline{\tilde{m}}_k^{(s_A)})^\top]^\top \ , \tag{6}$$

with $\underline{\tilde{m}}_k^{(s_i)} \in \mathcal{M}_k'$, and the measurement equation is given by

$$\underbrace{\begin{bmatrix} \underline{m}_k^{(a_1)} \\ \vdots \\ \underline{m}_k^{(a_A)} \end{bmatrix}}_{=:\underline{m}_k} = \underbrace{\begin{bmatrix} \underline{h}^{(a_1)}(\underline{x}_k) \\ \vdots \\ \underline{h}^{(a_A)}(\underline{x}_k) \end{bmatrix}}_{=:\underline{h}_k(\underline{x}_k)} + \underbrace{\begin{bmatrix} \underline{v}_k^{(a_1)} \\ \vdots \\ \underline{v}_k^{(a_A)} \end{bmatrix}}_{=:\underline{v}_k} \ , \tag{7}$$

where the zero-mean Gaussian measurement noise vector $\underline{v}_k$ has the covariance matrix $\mathbf{R}_k = \operatorname{diag}(\mathbf{R}_k^{(a_1)}, \dots, \mathbf{R}_k^{(a_A)})$. Therefore, the measurement $\underline{\tilde{m}}_k^{(s_i)}$ is a *realization* of the random vector $\underline{m}_k^{(a_i)}$. Note also that, if we receive less measurements than markers, not all markers are used during a measurement update to correct the state estimate, and thus, the human pose.

Finally, with the measurement (6) and the measurement model (7), we can directly apply the smart sampling Kalman filter (S^{2}KF) to update the predicted state estimate to obtain the posterior state mean $\hat{\underline{x}}_k^e$ and posterior state covariance $\mathbf{C}_k^e$.

F. Filter Initialization

Last but not least, to start with the recursive state estimation, an initial state estimate with initial state mean $\hat{\underline{x}}_0^e$ and initial state covariance $\mathbf{C}_0^e$ matrix is required. The estimator initialization strongly depends on the kinematic model, e.g., number of joints, and the utilized motion capture system. The initialization of our implementation will be discussed in Section IV.

G. Multiple Hypotheses Tracking (MHT)

In principle, a single Kalman filter would be sufficient to perform the whole-body motion tracking. Nonetheless, a main challenge in multi-target tracking with unknown associations is that the filter may converge to wrong local minima from which it cannot simply recover, i.e., in our case the filter would not converge to the true pose. Without forcing a special initialization pose with a specific root orientation and configuration of each extremity, e.g., the well-known T-pose, it is impossible to provide a single initial state estimate which lets the filter always converge to the correct pose.

To make our whole-body motion tracking more user-friendly and circumvent such a special initialization pose, we pursue a multiple hypotheses tracking (MHT) approach instead. The key idea of MHT is to maintain a tree of hypotheses to resolve the ambiguities in the state estimation arising from the unknown associations over time [6, Sec. 6.7]. However, unlike true MHT approaches and similar to [21], we do not form new hypotheses at each time step. Instead, we only generate multiple hypotheses at the very beginning when our initial pose is still entirely unknown and there is a significant risk of getting stuck in an incorrect pose. Note that this proposed setup is similar to [22] and is also a special case of an interacting multiple model (IMM) [23], as each filter has its own individual measurement model. However, the transition probability between different models is zero.

In essence, at time step k we have L_k filters with respective weights $w_k^{(l)}$, $1 \leq l \leq L_K$. How many filters are used in the beginning and how their respective initial state means and initial state covariances are determined will be discussed in Section IV. The filter weights form a discrete probability distribution over all filters and the overall pose estimate of the whole-body motion tracking is set to the estimate of the filter with the largest weight, i.e., the mode of the discrete probability distribution.

In each time step k, each filter performs a prediction based on the system model (4). It then computes the measurement (6) based on $\mathcal{M}_k$ and performs an update with the measurement equation (7). Subsequently, the current filter weights $w_k^{(l)}$ have to be updated for the next time step. Unfortunately, state-of-the-art weighting schemes such as evaluating the measurement in the measurement distribution [22] do not work due to the large measurement vector $\underline{m}_k$, as this leads to numerical issues. Hence, we again compute the optimal marker–measurement associations, but now with the already *updated* state estimate. As a by-product, we obtain the minimized sum $c_k^{(l)}$ of their Mahalanobis distances. Based on the $c_k^{(l)}$, the filter weights for the next time step are computed according to $w_{k+1}^{(l)} = w_k^{(l)}(c_k^{(l)})^{-1}$, $\forall l \in \{1, \dots, L_k\}$. The idea behind this is that filters that converge to a wrong pose will have more marker–measurement associations with larger Mahalanobis distances. As a result, filters with a small distance sum $c_k^{(l)}$ become more likely. Finally, the new filter weights have to be renormalized.

Over time, hypotheses become unlikely. To save computation time, we discard hypotheses that are no longer necessary until only a single hypothesis is left[2]. In order to discard hypotheses, we make use of the so-called effective sample size (ESS). The ESS is a prominent measure in the field of particle filtering, where the probability distribution of the system state is described by a set of weighted particles (instead of only a mean vector and a covariance matrix as in Kalman filtering). The idea of the ESS is to get information about the degeneracy of the particle set, i.e., how many particles have a weight close to zero. According to [3], for a set of P particles with *normalized* weights $\alpha^{(p)}$, the ESS is

$$\alpha_{\text{ESS}} = \frac{1}{\sum_{p=1}^{P} (\alpha^{(p)})^2} \ . \tag{8}$$

[2] Removing unlikely hypotheses and fusing similar hypotheses are two of the original pruning techniques proposed by Reid [24].

Due to the normalized weights, it holds that $1 \leq \alpha_{\mathrm{ESS}} \leq P$. For the extreme case that all particles are equally weighted, that is, no degeneracy, we have $\alpha_{\mathrm{ESS}} = P$. For the other extreme case that only a single particle has a non-zero weight, we have $\alpha_{\mathrm{ESS}} = 1$. Here, we compute (8) with the normalized filter weights $w_{k+1}^{(l)}$ and calculate the number of filters to be removed in this time step according to $R_k = \lfloor L_k - \alpha_{\mathrm{ESS}} + \epsilon \rfloor$. Then, the R_k filters with the *smallest* weights are removed, and thus, we also have $L_{k+1} = L_k - R_k$. Finally, we again have to renormalize the remaining filter weights.

The utilized rounding scheme with $0 \leq \epsilon < 1$ is necessary to effectively control when the last superfluous filter is removed if only two filters are left. Note that for $\epsilon = 0.5$ we have the usual rounding functionality. If ϵ would be zero, the last filter could only be removed when its weight becomes exactly zero. As this would require many time steps, we set $\epsilon = 0.05$. This means that if $L_k = 2$, the last filter will be eliminated when its weight drops below 0.5%.

However, it may happen that multiple filters converge to the true human pose. Consequently, their marker–measurement associations and Mahalanobis distances become very similar. As a result, their weights converge to nearly the same non-zero value, and thus, none of these filters will be removed by the procedure described above, although their information is redundant. Hence, we have to check if two filters represent nearly the same human pose. If so, the filter with the smaller weight gets removed. We check for similarity based on three indicators: (i) the Euclidean distance in the root pose is smaller than a threshold ε_p, (ii) the largest root orientation difference is smaller than a threshold ε_r, and (iii) the largest joint angle difference is smaller than a threshold ε_a.

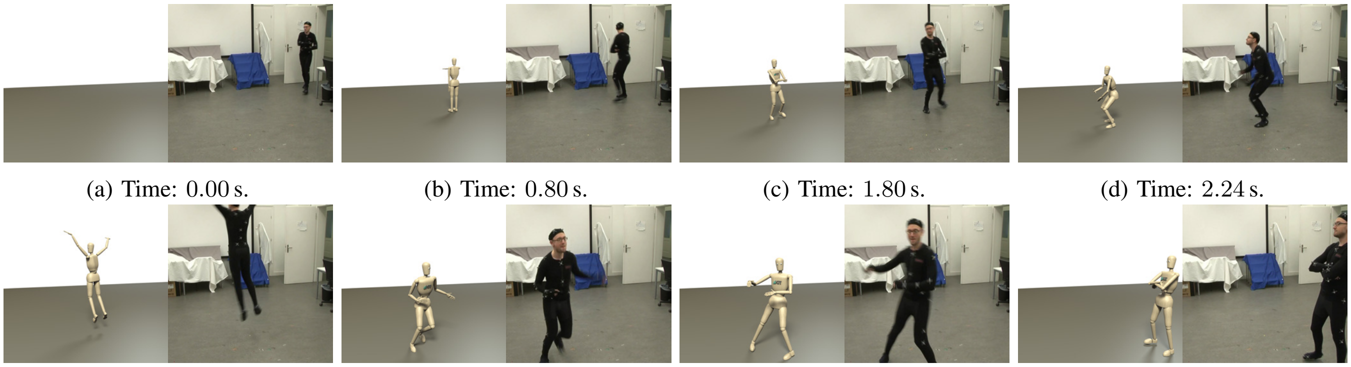
IV. Implementation for the MMM Model and a Vicon Optical Motion Capture System

In this section, we describe a concrete implementation of our proposed approach for whole-body motion tracking. We rely on the MMM reference model presented in Section II (scaled to the body height of the human to be tracked) with its kinematic model for the human pose, including the placement of $N = 53$ markers. In total, $J = 48$ joint angles are used for the kinematic model (eyes and fingers are excluded), resulting in a system state dimension of $D = 54$. Moreover, root position and marker positions are measured in millimeters and root orientation and joint angles are measured in radians (this is important as it also defines the units of the noise covariance matrices $\mathbf{R}_k^{(n)}$ and $\mathbf{Q}_k$). The marker positions are measured by a Vicon MX10 system using ten T10 cameras, which is an optical motion capture system based on passive (reflective) markers. The system captures at 100 Hz, that is, every 10 ms, we get a new set of markers $\mathcal{M}_k$. For the measurement update, the measurement noise properties of the Vicon system, i.e., the covariance matrices $\mathbf{R}_k^{(n)}$, have to be known. Experimentally, we have found that the marker positions provided by the Vicon system are approximately disturbed by a Gaussian noise with a covariance of $\mathbf{R}_k^{(n)} = 10^{-3} \cdot \mathbf{I}_3$, $\forall 1 \leq n \leq N$, where $\mathbf{I}_3$ denotes the identity matrix of dimension three. To perform the measurement update, the S^{2}KF is configured to use $S = 351$ samples. Furthermore, the system noise covariance matrix is set to the time-invariant diagonal matrix $\mathbf{Q}_k = \mathrm{diag}(25 \cdot \mathbf{I}_3, 10^{-10} \cdot \mathbf{I}_3, 10^{-9} \cdot \mathbf{I}_{48})$. The threshold for measurement outliers is set to $\varepsilon_o = 300\,\mathrm{mm}$ and the thresholds for pose similarity to $\varepsilon_p = 1\,\mathrm{mm}$, $\varepsilon_r = 0.001\,\mathrm{rad}$, and $\varepsilon_a = 0.01\,\mathrm{rad}$, respectively.

Regarding the initialization of the tracking, on the one hand, we have to keep the number of initial poses, i.e., filters, as small as possible to be able to operate in real-time. On the other hand, we have to cover as many different initial poses as possible to maximize the probability of converging to the true human pose. In order to get an adequate number, our only restriction on the initial pose is that the human subject is standing upright. Consequently, the initial state means for joint angles of the legs and torso are the same for all initial poses, and the focus is on possible arm configurations. Here, we select five significantly different configurations per arm and build the Cartesian product for both arms, which results in 25 different poses. Moreover, to cover different root orientations, each of these 25 poses is rotated in 90 degree steps, i.e., the initial yaw angle means are set to $\hat{o}_0^y \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ and the roll and pitch angle means are $\hat{o}_0^r = \hat{o}_0^p = 0$ for all poses. This leads to a total number of $L_0 = 100$ initial pose hypotheses/filters. Furthermore, the initial mean $\hat{\underline{r}}_0$ and covariance $\mathbf{C}_0^r$ of the root position for all 100 filters are set to sample mean and sample covariance of the first set of available marker measurements $\mathcal{M}_0$. Last but not least, the initial covariance matrix for the root orientation is set to $\mathbf{C}_0^o = 10^{-6} \cdot \mathbf{I}_3$, whereas the initial covariance matrix of the joint angle parameters is set to $\mathbf{C}_0^\Theta = 10^{-10} \cdot \mathbf{I}_{48}$. Hence, the overall initial state covariance matrix for all filters is given by $\mathbf{C}_0^e = \mathrm{diag}(\mathbf{C}_0^r, \mathbf{C}_0^o, \mathbf{C}_0^\Theta)$.

When the tracking starts, all 100 filters have to process the incoming measurements. Unfortunately, doing this in real-time is hardly possible. Nonetheless, the subsequent sets of measurements are not discarded. Instead, they are queued up for later processing. If we omitted these measurements, the result would be a track loss due to potential comprehensive changes in the human pose and the long time between two consecutive updates. Consequently, at the beginning of the tracking, we accept a substantial and non-negligible lag. However, as evaluations in Section V show, the number of active filters rapidly decreases and one or two remaining filters require less than 10 ms. Thus, the tracking approach can compensate the lag in a short time to finally achieve the desired real-time capability when only few filters remain.

In summary, due to the avoidance of a special initialization pose, we are able to offer a user-friendly tracking, as the subject can simply walk into the area being observed by the Vicon system and tracking automatically starts if a sufficient number of marker measurements becomes available. In order to have a sufficient number of markers for the initialization, we configure the tracking to start when it receives at least 48 marker measurements for the first time. After that, there is no restriction on the number of measurements to be processed.



(a) Time: 0.00 s.

(b) Time: 0.80 s.

(c) Time: 1.80 s.

(d) Time: 2.24 s.

(e) Time: 2.68 s.

(f) Time: 3.56 s.

(g) Time: 3.92 s.

(h) Time: 5.36 s.

Fig. 2: A sequence of fast and complex whole-body motions performed by a subject and their pose tracked by our proposed approach. The images on the right show the performed motions from a time-synchronized video recording, while the images on the left indicate the estimated pose.

V. Evaluation

In this section, we evaluate the implementation of the proposed whole-body motion tracking from Section IV. Since a ground truth, i.e., the true system state, cannot be obtained for captured human motions, usual state estimation metrics such as the normalized estimation error squared (NEES) are not applicable. Therefore, we studied the convergence and runtime performance for 20 observed human motions. In 18 of these motions, the tracking attains the correct pose and also challenging initial poses, such as walking backwards into the area observed by the motion capture system, pose no problems.

As an example, we demonstrate a reconstructed sequence of fast and complex motions in Fig. 2. At the beginning, the subject walks into the area observed by the Vicon motion capture system (see Fig. 2a), and the tracking has not recognized the user yet. After about 0.8 s, the tracking automatically starts as it obtains 48 marker measurements for the first time (see Fig. 2b). Note that the subject does not exhibit the T-pose required by the approach presented in [9], but instead shows a challenging initial pose with the arms being folded very close to the body. It can be seen that after the first measurement update, the pose estimate already has an approximately correct root pose and root orientation. After another second, the tracking has already converged to the correct pose (Fig. 2c). Over time, fast movements of the body and the arms are performed. This includes several fast turns, e.g., from Fig. 2c to Fig. 2d, a jump with outstretched arms (Fig. 2e), another turn (Fig. 2f), fast moving arms (Fig. 2g), and again a folding of the arms at the end (Fig. 2h). In conclusion, the good tracking results indicate that the measurements are correctly associated most of the time.

Fig. 3a shows the varying number of received marker measurements over time. At the beginning, the subject walks into the room and more and more markers become visible to the Vicon system. After 1.8 s, there is a jump in the number of measurements. This is due to the unfolding of the subject's arms as now more of the markers attached to arms and hands can be measured by the Vicon system. It should be noted that between 2.5 s and 3 s, we get more than 53 measurements, and thus, definitively have clutter measurements. Of course, clutter may also be present at other time steps. After 4.8 s, the subject folds their arms again, which decreases the number of measurements (see Fig. 2h).

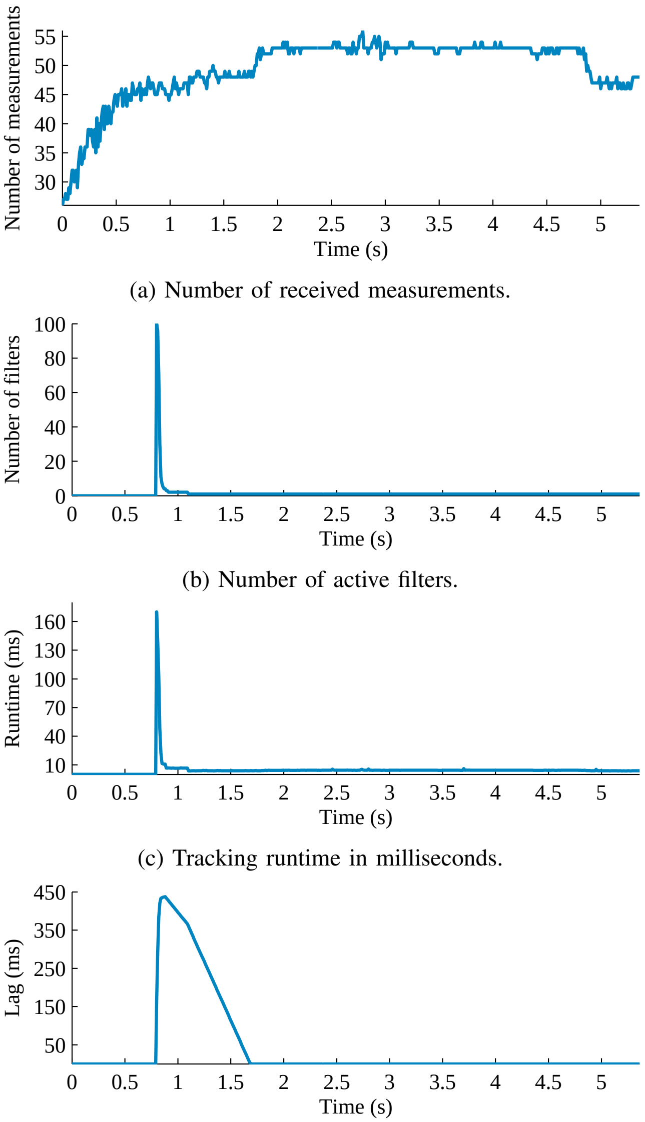
Fig. 3b depicts the number of active filters used by the tracking. Prior to 0.8 s, the tracking is inactive and no filter is in use. At the beginning of the tracking, all initial 100 filters become active. However, the number of active filters drops considerably fast. At 1.1 s, only a single filter is left. The number of active filters massively affects the runtime of the motion tracking (see Fig. 3c). With 100 active filters, we have a peak runtime of 170 ms. However, if only a single filter is left, we measure a runtime of 4 ms on average, resulting in real-time tracking.

As already mentioned in Section IV, the initially large runtimes cause a significant lag in the processing of the received measurements. This is shown in Fig. 3d. The lag peaks at about 440 ms. That is, the estimation result based on the first measurements received at 0.8 s is actually available at 1.24 s. Nevertheless, after the number of active filters drops significantly, it takes less than 10 ms to update the estimate and the lag decreases in a linear fashion. At 1.7 s, the lag has already been completely compensated and we get the tracking results in real time. In conclusion, it takes only 900 ms to process measurements in real time after the tracking started.

In general, in our 20 evaluated motions, all filters except one have been eliminated after a maximum of 1.9 s, and real-time capabilities are attained in a similar way to the exemplary motion.

VI. Conclusions

In this paper, we presented a novel approach to track whole-body human motions with unlabeled marker measurements in real time. The approach is based on four key components: (i) a known kinematic model of the human body that includes the locations of the attached markers, (ii) constrained sample-based Kalman filtering, (iii) the LAPJV algorithm, a fast version of the Hungarian algorithm, to obtain optimal marker–measurement associations, and (iv) a multiple hypotheses tracking approach to avoid a special initialization pose and



(a) Number of received measurements.

(b) Number of active filters.

(c) Tracking runtime in milliseconds.

(d) Tracking lag in milliseconds.

Fig. 3: Details of the tracking for the motion shown in Fig. 2. The evaluation was performed on an Intel Core i7-3770 (3.4 GHz).

convergence to a wrong pose. We implemented the proposed tracking approach for the MMM reference model and a Vicon optical marker-based motion capture system, and evaluated our implementation in various scenarios. From these experiments, we conclude that the proposed whole-body motion tracking can accurately estimate the human pose over time and can handle unavoidable marker occlusions or clutter measurements easily. Although the MHT approach leads to substantial time lag at the beginning, it can be compensated very fast, usually in about a second, and after that the tracking can operate in real time.

References

[1] Dan Simon, *Optimal State Estimation*, 1st ed. Wiley & Sons, 2006.

[2] Pawe Stano, Zsófia Lendek, Jelmer Braaksma, Robert Babuska, Cees de Keizer, and Arnold J. den Dekker, "Parametric Bayesian Filters for Nonlinear Stochastic Dynamical Systems: A Survey," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1607–1624, Dec. 2013.

[3] Branko Ristic, Sanjeev Arulampalam, and Neil Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Publishers, 2004.

[4] Jannik Steinbring, Christian Mandery, Nikolaus Vahrenkamp, Tamim Asfour, and Uwe D. Hanebeck, "High-Accuracy Real-Time Whole-Body Human Motion Tracking Based on Constrained Nonlinear Kalman Filtering," *arXiv preprint: Systems and Control (cs.SY)*, Nov. 2015.

[5] Jannik Steinbring, Martin Pander, and Uwe D. Hanebeck, "The Smart Sampling Kalman Filter with Symmetric Samples," *Journal of Advances in Information Fusion*, vol. 11, no. 1, pp. 71–90, Jun. 2016.

[6] Samuel Blackman and Robert Popoli, *Design and Analysis of Modern Tracking Systems*. Artech House Publishers, Jul. 1999.

[7] Ronald P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Artech House, Feb. 2007.

[8] Jonathan Maycock, Tobias Röhlig, Matthias Schröder, Mario Botsch, and Helge Ritter, "Fully Automatic Optical Motion Tracking using an Inverse Kinematics Approach," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, Seoul, South Korea, Nov. 2015, pp. 461–466.

[9] Johannes Meyer, Markus Kuderer, Jörg Müller, and Wolfram Burgard, "Online Marker Labeling for Fully Automatic Skeleton Tracking in Optical Motion Capture," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, May 2014, pp. 5652–5657.

[10] Tobias Schubert, Alexis Gkogkidis, Tonio Ball, and Wolfram Burgard, "Automatic Initialization for Skeleton Tracking in Optical Motion Capture," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, USA, May 2015, pp. 734–739.

[11] Andreas Aristidou and Joan Lasenby, "Real-Time Marker Prediction and CoR Estimation in Optical Motion Capture," *The Visual Computer*, vol. 29, no. 1, pp. 7–26, 2013.

[12] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour, "Unifying Representations and Large-Scale Whole-Body Motion Databases for Studying Human Motion (to appear)," *IEEE Transactions on Robotics*, 2016.

[13] Ömer Terlemez, Stefan Ulbrich, Christian Mandery, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour, "Master Motor Map (MMM) – Framework and Toolkit for Capturing, Representing, and Reproducing Human Motion on Humanoid Robots," in *2014 14th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, Madrid, Spain, Nov. 2014, pp. 894–901.

[14] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour, "The KIT Whole-Body Human Motion Database," in *17th International Conference on Advanced Robotics (ICAR 2015)*, Istanbul, Turkey, Jul. 2015, pp. 329–336.

[15] Christian Mandery, Júlia Borràs, Mirjam Jöchner, and Tamim Asfour, "Analyzing Whole-Body Pose Transitions in Multi-Contact Motions," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, Seoul, South Korea, Nov. 2015, pp. 1020–1027.

[16] Ondřej Straka, Jindřich Duník, and Miroslav Šimandl, "Truncation Nonlinear Filters for State Estimation with Nonlinear Inequality Constraints," *Automatica*, vol. 48, no. 2, pp. 273–286, Feb. 2012.

[17] Martin Liggins, David Hall, and James Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice*, 2nd ed. CRC Press, Sep. 2008.

[18] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[19] R. Jonker and A. Volgenant, "A Shortest Augmenting Path Algorithm for Dense and Sparse Linear Assignment Problems," *Computing*, vol. 38, no. 4, pp. 325–340, Dec. 1987.

[20] Dimitri P. Bertsekas, "Auction Algorithms," in *Encyclopedia of Optimization*. Springer US, 2001, pp. 73–77.

[21] Florian Faion, Marcus Baum, Antonio Zea, and Uwe D. Hanebeck, "Depth Sensor Calibration by Tracking an Extended Object," in *Proceedings of the 2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2015)*, San Diego, USA, Sep. 2015, pp. 19–24.

[22] N. Peach, "Bearings-only Tracking Using a Set of Range-parameterised Extended Kalman Filters," *IEE Proceedings of Control Theory and Applications*, vol. 142, no. 1, pp. 73–80, Jan. 1995.

[23] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan, "Interacting Multiple Model Methods in Target Tracking: A Survey," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 1, pp. 103–123, Jan. 1998.

[24] Donald B. Reid, "An Algorithm for Tracking Multiple Targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.