A stochastic algorithm for automatic hand pose and motion estimation

Francesca Cordella, Francesco Di Corato, Bruno Siciliano & Loredana Zollo

Medical & Biological Engineering & Computing

ISSN 0140-0118

Med Biol Eng Comput
DOI 10.1007/s11517-017-1654-6

Springer

Med Biol Eng Comput
DOI 10.1007/s11517-017-1654-6

ORIGINAL ARTICLE

A stochastic algorithm for automatic hand pose and motion estimation

Francesca Cordella[1] · **Francesco Di Corato**[3] · **Bruno Siciliano**[2] · **Loredana Zollo**[1]

Received: 23 September 2016 / Accepted: 27 April 2017

Abstract In this paper, a novel, robust, and simple method for automatically estimating the hand pose is proposed and validated. The method uses a multi-camera optoelectronic system and a model-based stochastic algorithm. The approach is marker-based and relies on an Unscented Kalman Filter. A hand kinematic model is introduced for constraining relative marker's positions and improving the algorithm robustness with respect to outliers and possible occlusions. The algorithm outputs are 3D coordinate measures of markers and hand joint angle values. To validate the proposed algorithm, a comparison with ground truths for angular and 3D coordinate measures is carried out. The comparative analysis shows the advantages of using the model-based stochastic algorithm with respect to standard processing software of optoelectronic cameras in terms of implementation simplicity, time consumption, and user effort. The accuracy is remarkable, with a difference of maximum $0.035rad$ and $4mm$ with respect to angular and 3D Cartesian coordinates ground truths, respectively.

Keywords Hand pose estimation · Unscented Kalman filter · Optoelectronic cameras · Hand motion analysis

1 Introduction

Hand motion analysis finds application in robotics, with reference to learning-by-demonstration approaches and grasping database creation [1], in virtual reality and medical fields.

Reconstructing the motion of the human hand joints is complex due to the high number of Degrees of Freedom (DoFs). Position, pressure, electromagnetic and inertial sensors [2] so as data-gloves [3] are some of the methods used for estimating the hand pose, but they suffer of several limitations (drift, obtrusiveness, customized calibration, etc.).

Vision-based tracking systems represent a valuable alternative to the above systems since they can be utilized with hands of different size and let the user perform more natural movements.

The accuracy of RGB-D cameras, recently adopted for tracking hands [4, 5], is still far from that obtained with optoelectronic cameras [6]. Therefore, although these latter are still expensive and require a completely structured environment for calibration and acquisition, they are widely used for motion analysis. However, presently used approaches [7] for hand motion analysis are very time-consuming and, more importantly, generally require that the association between physical markers and measurements is known. The software provided by manufacturers for

Work done while Francesco Di Corato was at the Research Center "E. Piaggio" Pisa, Italy

✉ Francesca Cordella
f.cordella@unicampus.it

Francesco Di Corato
dicorato.francesco@gmail.com

Bruno Siciliano
bruno.siciliano@unina.it

Loredana Zollo
l.zollo@unicampus.it

[1] Unit of Biomedical Robotics and Biomicrosystems, Università Campus Bio-Medico di Roma, via Alvaro del Portillo 21, 00128, Rome, Italy

[2] PRISMA Lab, Department of Electrical Engineering and Information Technology, Università di Napoli Federico II, via Claudio 21, 80125, Naples, Italy

[3] Research Center "E. Piaggio", Pisa, Italy

Published online: 08 June 2017

labeling and tracking markers over time is unable to solve all the associations, requiring the manual intervention of the user, thus making the data user-dependent. Furthermore, these systems require post-processing for extracting information about hand joint angles.

Methods for online marker labeling have been also applied in human motion analysis, e.g., [8] and [9], but are applicable only for skeleton tracking.

In [10], a model-based approach has been adopted for hand tracking. The models have been uploaded from a database created ad hoc and the computational cost, although not reported, seems to be quite high being necessary a preprocessing phase. The authors emphasize the use of a reduced number of markers, but this increases the error on the joint angle estimation in the case of the human hand. In [11], an Unscented Kalman Filter (UKF) is implemented for predicting marker position. The main limitation is that a hand model is not adopted implying the possibility to have not realistic hand configurations.

The method illustrated in this paper aims at overcoming the abovementioned limitations by proposing an automatic, robust, and self-contained stochastic algorithm for hand pose and motion estimation that is grounded on an UKF [12] and employs reflective markers and an optoelectronic multicamera system. The proposed technique is able to adaptively associate a given image measurement to a marker or to an outlier by using probabilistic techniques. The algorithm reformulates the problem into a stochastic nonlinear filtering framework through UKF and relies on an appropriate hand kinematic model whose accuracy has been validated. With respect to other nonlinear estimation techniques, such as Extended Kalman Filters (EKF), the UKF has proven to improve the estimation performance, is very simple to be implemented thanks to the software modularity, and does not need the computation of Jacobian matrices, as required by the EKF.

Differently from approaches like [8] and [9], the present one is general, being applicable to any object whose geometry is defined in an online initialization phase. Creating the hand model in the initialization phase avoids the use of a clutter database, like in [10], making the approach applicable to different hands. Furthermore, the proposed method has a low computational burden, by paving the way for a real-time implementation, as discussed in Section 3.3. Adopting a marker on each hand joint guarantees a remarkable accuracy, as shown in Section 3.4.

The approach has been validated by means of a comparative analysis between the estimated joint angle values and an angular ground truth, made of the measures obtained from the joint angle sensors embedded in a robotic hand (i.e., the DLR-HIT Hand II [13]) and between the estimated 3D Cartesian coordinates of the finger joints and a 3D Cartesian coordinates ground truth, made of the marker 3D Cartesian coordinates obtained from an optoelectronic system (i.e., the BTS Smart D www.btsbioengineering.com/it/). To this purpose, the optoelectronic system has been used to track markers positioned both on the subject and on the robotic hands. Although this paper is focused on the hand motion estimation, the proposed approach is general. It means that, once known the kinematic parameters of the object to be tracked, the method automatically reconstruct the motion and pose.

2 Methods

2.1 Hand kinematic model

An accurate kinematic model is needed for mapping sensor information to coordinate frames and joint angles. In order to describe human hand kinematics, different kinematic models have been proposed [14]. They differ for simplifying assumptions, especially related to the number of DoFs and the position and orientation of the Axis of Rotation (AoR) [15, 16].

In this work, a 21-DoFs hand model is considered [17]: the thumb is modeled as a 5-DoFs kinematic chain, whereas the long fingers have 4 DoFs each. Joint angle values have been determined after placing 25 reflective markers of 6-mm diameter on the right hand of a volunteer subject, as shown in Fig. 1. The protocol for positioning markers on the hand has been chosen in order to minimize artifacts (due for instance to skin movements or marker occlusion) and to obtain information about wrist position. Four markers have been positioned on each finger in correspondence of MetaCarpo-Phalangeal (MCP), Proximal Inter-Phalangeal (PIP), Distal Inter-Phalangeal (DIP) joints, and fingertips (TIP). There is one more marker on the thumb, positioned close to the TrapezioMetacarpal (TM) joint (in Fig. 1 it is called TMb): this simplifies the determination of the AoR of the TM joint. Three markers (called $B1$, $B2$, $B3$) positioned on the hand dorsum constitute the system reference frame: the dorsum is the part of the hand that suffers less from skin movements.

The BTS optoelectronic system gives the 3D Cartesian coordinates of the center of each marker. It has been assumed that the Cartesian coordinates of the marker centers correspond to the finger joints Cartesian coordinates and that the Cartesian coordinates of the finger joints reference frame origin is coincident with the corresponding finger joint Cartesian coordinates.

The axes of rotation of the long finger MCP joints and of the thumb TM joint are defined as shown in Fig. 1; further, it has been supposed that the Flexion/Extension (F/E) axes of PIP and DIP joints are parallel to each other.

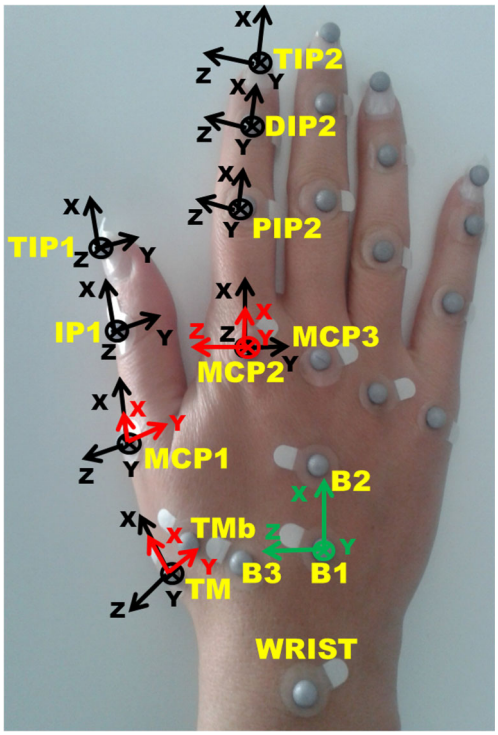


Fig. 1 Joint reference frames and adopted protocol for marker positioning. Joint labels are outlined in yellow. The reference frame, positioned on the hand palm and outlined in green, has X-axis along the line connecting marker $B1$ with marker $B2$, Y-axis perpendicular to the palm plane and Z-axis defined with the right hand rule

The hand reference frame (outlined in green in Fig. 1) is assumed to be centered in marker $B1$ and its orientation is shown in the figure. It represents the base frame for the hand body, with respect to which the forward kinematics of the hand is expressed. The hand palm is modeled as a rigid body with 6 DoFs, consisting of 3 components of translation and 3 angles of rotation. The angular parametrization of Euler angles in configuration ZYX has been chosen in order to obtain the hand joint angles from the rotation matrices [18] (see the Appendix for formulation of the equations).

2.2 Stochastic hand motion and pose estimation

A robust pose estimation scheme able to estimate the relative motion —in terms of position, orientation and velocity— of a tracked hand (whose kinematics is known) with respect to a multi-camera optoelectronic system is described in this section.

2.2.1 Modeling and filtering hand motion and pose

A multi-camera optoelectronic system is adopted in order to obtain information about the motion and the pose of the hand in the operating space. The noisy measurements of the $3D$ positions of the markers placed onto the hand surface with respect to the camera are thus available from the vision system.

An ideal vision algorithm, capable of maintaining coherent correspondences between measurements and physical markers, would express the measurement of the N markers placed onto the hand surface as

$$y(t) = \begin{bmatrix} y_i(t) \\ \vdots \\ y_l(t) \end{bmatrix} = \begin{bmatrix} g(t)T_{m_i b} + v_i(t) \\ \vdots \\ g(t)T_{m_l b}(\Theta) + v_l(t) \end{bmatrix}. \quad (1)$$

where, $y_i(t)$ is the measurement of the i-th marker on the hand dorsum, $y_l(t)$ is the measurement of the l-th marker on the finger joints, $v_i(t) \sim N(0, R)$ and $v_l(t) \sim N(0, R)$ represent white zero-mean, normally distributed stochastic processes, with variance assumed constant among features. The position of the l-th marker with respect to the hand reference frame is expressed in terms of the generalized vector of joint variables $\Theta = \left[\dots \ \theta_k^r(t) \ \dots \right]^T$ of the fingers and is obtained via forward kinematics, i.e. $T_{m_l b}(\Theta)$. $\theta_k^r(t)$ denotes the k-th joint angle of the r-th finger.

The motion of the hand reference frame b (whose origin is rigidly attached to marker $B1$) with respect to the camera frame c, assumed fixed in space, is modeled according to the following discrete-time model

$$\begin{cases} T(t+1) = T(t) + v(t)dt \\ v(t+1) = v(t) + \eta_v(t)dt \\ R(t+1) = R(t)e^{\Omega(t)dt} \\ \theta_k^r(t+1) = \theta_k^r(t) + \eta_{\theta_k^r}(t)dt \\ \Omega(t) = \eta_\omega(t)\wedge \\ y_i(t) = T_{m_i b}g(t) + v_i(t), \ i \in \mathcal{V}_i(t) \subseteq \{1, 2, ..., N_p\} \\ y_l(t) = T_{m_l b}(\Theta)g(t) + v_l(t), \ l \in \mathcal{V}_l(t) \subseteq \{1, 2, ..., N_f\} \end{cases} \quad (2)$$

where $T(t)$, $v(t)$ and $R(T)$ are position, linear velocity and rotation matrix of the hand frame with respect to the camera frame, respectively, $\Omega(t)$ is the skew-symmetric matrix of angular velocity $\eta_\omega(t)$ expressed in the coordinates of the hand frame (being $\wedge$ the cross-product operator). Variables $T(t)$ and $R(t)$ define the group transformation $g(t) \triangleq \{R(t), T(t)\} \in SE(3)$ that fully describe the 6-DoF localization problem of the hand with respect to the defined reference frame fixed in space. $\mathcal{V}_i(t)$ and $\mathcal{V}_l(t)$ denote the group of markers (on the hand dorsum and the fingers) that are visible at the current time (except for the clutters). Therefore, the sets $\mathcal{V}_i(t)$ and $\mathcal{V}_l(t)$ are time dependent, since the physical markers could be not visible due to illumination artifacts or occlusions. It is assumed that no prior information regarding the nature of the hand motion and the time evolution of the joint angles is available. Hence, zero-mean

white Gaussian noises with constant variance $\eta_v(t)$, $\eta_\omega(t)$, $\eta_{\theta_k^r}(t)$ are introduced. They model the hand linear accelerations, the angular velocities, and the joint angular velocities as random walks.

As in all pose estimation algorithms, the rationale behind the proposed approach is to use the measurements obtained from the visual system as a measure of the pose and motion variables of the hand. This measure is then used to estimate the state variables in Eq. 2. The novelty of the approach, in the framework of hand pose estimation, is twofold: (i) The marker measurements are assumed to be randomly provided. This means that the association between a measurement and the right entity (outlier or physical marker) cannot be made a priori, but is made adaptively. (ii) The introduction of the information about the hand anatomy in the tracking algorithm improves its robustness with respect to outliers, occlusions, and markers entering and exiting through the field of view. In fact, as evident from Eq. 2, the measurements of the i-th and of the l-th markers are given by the forward kinematics. It means that the marker measurements are constrained by each other: their motion should respect the constrains imposed by the kinematic models. If a marker disappears, its position can be estimated on the basis of the nearest marker position and of the kinematic model.

According to the motion and pose dynamics in Eq. 2, given the visual measurements of the markers, Eq. 1, (with $N = 25$), a nonlinear stochastic estimation scheme has been designed. The aim of the filter is to estimate the state $x(t)$ of the system, consisting of the motion variables, $T(t), v(t)$, the angular parametrization of the rotation matrix $R(t)$, and the value of the finger joint angles.

Given the strong non-linearity of the model with respect to the state and the noise terms, a certain number of estimation schemes can be implemented (such as EKF, UKF, and particle filters). Although other choices can be made, this work is not focused on estimating the whole conditional density function of the state, as in particle evolution schemes, but only the point estimate of the state, since a unimodal posterior density of the state is expected. All the deviations from the nominal model assumptions (i.e., the tails of the posterior) are considered to be due to clutters, and are ignored in the estimation process. This fact, together with the Gaussian nature of the noises involved into state evolution and measurements, motivated to limit the discussion to nonlinear Kalman filtering. Thus, an estimation scheme based on an augmented Unscented Kalman Filter (UKF) [12] has been selected. The peculiarity of the adopted estimation scheme, as compared with the classical UKF approach [19], is the possibility to easily deal with non-affine noise terms in the state/measurement model.

2.2.2 Robust marker tracking – the association problem

The tracking phase in passive marker-based visual systems may be problematic due to local illumination changes, reflections or occlusions. Since a model of the body and the information about its shape are available, it has been chosen to reformulate the tracking problem into a stochastic optimization problem, embedded into the estimation task.

To this aim, the raw outputs given by the visual system algorithm, corresponding to the image at time t, are considered as a random sequence of M_t measurements $\mathbf{y}_t = \{y_1(t), y_2(t), ..., y_{M_t}(t)\}$ of marker candidates. In general, condition $M_t \neq N$ holds, which means that the sequence $\mathbf{y}_t$ does contain measurements of visible markers and outliers.

The randomness of the measurement sequences is a fundamental issue in this framework, since it implies the following important consequences: (i) the associations between measurement y_i and marker j or with a clutter cannot be decided a priori and has to be estimated; (ii) each sequence of measurements for each frame can be considered conditionally independent on every other sequence in the past; (iii) once the current sequence of associations has been defined, it can be considered conditionally independent on the past history of associations. A direct consequence is that predicting the order in which markers and clutters are detected, for each image, can be very tricky. Because of the above assumption, a solution to the filtering problem (ensuring robustness at the same time) has been developed by employing probabilistic techniques. In order to model the measurement-to-marker or measurement-to-clutter association [20], a latent variable $a_i(t)$ has been introduced for each measurement $y_i(t) \in \mathbf{y}_t$.

$$a_i(t) = \begin{cases} 0, & \text{if } y_i(t) \text{ is a clutter} \\ j, & \text{if } y_i(t) \text{ is the measurement of marker } j. \end{cases} \quad (3)$$

The compact form of the nonlinear model (1), i.e., $y(t) = h(x(t))$ (where $x(t)$ is the state variable), can be considered as a conditional measurement model over the variable $a_i(t)$. In fact, the output function can be written as $y_i(t) = h(x(t)\,|a_i(t) = j \neq 0)$, being conditioned over a certain value of the latent variable. It means that the rows corresponding to the measurement of the marker j can be selected from the function $h(x(t))$. If $a_i(t) = 0$, the output model becomes $y_i = v_o$, $v_o \sim \mathcal{N}(\bar{v}_o, \Sigma_o)$. Therefore, the association problem consists in maximizing the belief that the current measurement $y_i(t) \in \mathbf{y}_t$ corresponds to either a visible marker or a clutter. It can be performed by finding the most probable value of the variable $a_i(t)$, $\forall i = 1, \ldots, M_t$, for every measurement collected at the current time step.

Formalizing, the aim is to find the maximum of the posterior distribution

$$p\left(a_i(t)\,|y_i(t), \mathbf{y}_{0:t-1}\right) \propto p(y_i(t)\,|a_i(t), \mathbf{y}_{0:t-1})\,p(a_i(t)) \tag{4}$$

given the current observation $y_i(t)$ and the whole history of the measurements up to the previous step. Eq. 4 has been obtained via the Bayes' rule. In Eq. 4:

- $p\left(a_i(t)\right)$ is assumed to be independent on the previous measurements and is determined by the a priori knowledge of clutter and marker association event probabilities. Since no specific prior probability distribution is generally available, one possible choice is to consider the probability of detecting the marker j as the same of detecting the marker $d \neq j$, i.e. by considering a uniform distribution for the marker association.
- $p\left(y_i\,|a_i, \mathbf{y}_{0:t-1}\right)$ is the likelihood that the current measurement is associated to a given marker or to a clutter and can be written as

$$p\left(y_i\,|a_i, \mathbf{y}_{0:t-1}\right) = \int p(\,y_i|\,x, a_i, \mathbf{y}_{t-1})\,p(x\,|a_i, \mathbf{y}_{t-1})\,dx \tag{5}$$

$$= \int p\left(y_i\,|x, a_i, \mathbf{y}_{t-1}\right)\,p\left(x\,|\mathbf{y}_{t-1}\right)\,dx \tag{6}$$

Upon fixing a certain guess for the association, $a_i(t) = j,\ j \neq 0$, Eq. 6 is the Kalman Filter likelihood of the measurement $y_i(t)$, given the prediction of the marker j, i.e. the conditioning of the measurement model over that value of the latent variable. Thus, given the predicted state-related Sigma-Points [12], $\mathbf{X}^x_{n,t/t-1},\ n = 1, \ldots, L$, computed by employing the nonlinear state model, their transformation through the conditioned measurement function can be obtained, as in a classical UKF

$$\mathbf{Y}^j_{n,t/t-1} = h\left(\mathbf{X}^x_{n,t/t-1}\,|a_i = j\right). \tag{7}$$

The superscript j on the transformed Sigma-Points of the output indicates that $Y^j_{n,t/t-1}$ refers to the predicted projection of the marker j, for which the association is being tested. The mean and the covariance of the measurement vector are calculated as

$$\widehat{y_j^-} = \sum_{n=0}^{L} W_m^n \mathbf{Y}^j_{n,t/t-1} \tag{8}$$

$$P^-_{yy,j} = \sum_{n=0}^{L} W_c^n \left(\mathbf{Y}^j_{n,t/t-1} - \widehat{y_j^-}\right)\left(\mathbf{Y}^j_{n,t/t-1} - \widehat{y_j^-}\right)^T + R_j \tag{9}$$

where W_c^n and W_m^n are the weights associated to the Sigma-Points [12], $\widehat{y_j^-}$ is the predicted projection of the marker j and $P^-_{yy,j}$ its covariance. Thus, the probability of the association $a_i = j$ can be computed as

$$p\left(a_i = j\,|y_i, \mathbf{y}_{0:t-1}\right) \propto \mathcal{N}\left(y_i - \widehat{y_j^-},\, P^-_{yy,j}\right)\,p(a_i = j) \tag{10}$$

being $\mathcal{N}()$ the multivariate normal distribution of proper mean value and covariance. The set of possible associations is discrete, and thus the (discrete) value of the association posterior distribution can be computed by inspecting all the possible values of the associations (per each measurement) as follows

$$\begin{cases} p\left(a_i = 0\,|y_i, \mathbf{y}_{0:t-1}\right) \propto \frac{1}{RES_u \times RES_v}\,p\left(a_i = 0\right) \\ p\left(a_i = 1\,|y_i, \mathbf{y}_{0:t-1}\right) \propto \mathcal{N}\left(y_i - \widehat{y_1^-},\, P^-_{yy,1}\right)\,p\left(a_i = 1\right) \\ \vdots \\ p\left(a_i = N\,|y_i, \mathbf{y}_{0:t-1}\right) \propto \mathcal{N}\left(y_i - \widehat{y_N^-},\, P^-_{yy,N}\right)\,p\left(a_i = N\right) \end{cases}. \tag{11}$$

In the case of clutter association ($a_i = 0$) the likelihood function has been set equal to $1/\left(RES_u \times RES_v\right)$, where $RES_u \times RES_v$ is the image resolution, meaning that a clutter can happen everywhere in the image. This choice is usually considered valid in approaches similar to the one proposed here, for example [21]. Selecting the maximum probability among the ones in Eq. 11, will give the most probable value of the variable $a_i(t)$, corresponding to the measurement $y_i(t)$. The association problem is solved by repeating the above procedure for all the measurements in the set $\mathbf{y}_t$.

In the following, the conditions needed to perform the Kalman correction step, given the solution to the association problem, are determined and how to perform such correction is explained. When Eq. 11 is applied to the entire measurement set, the sequence of probabilities can be normalized and put into a matrix called *Feasible Association Matrix*. It can be expressed as

$$\mathcal{F}_{M_t} = \begin{bmatrix} \pi_{10} & \pi_{11} & \ldots & \pi_{1N} \\ \pi_{20} & \pi_{21} & \ldots & \pi_{2N} \\ \vdots & & \ddots & \vdots \\ \pi_{M_t 0} & \pi_{M_t 1} & \ldots & \pi_{M_t N} \end{bmatrix} \tag{12}$$

where $\pi_{ij} = \frac{p\left(a_i = j\,|y_i, \mathbf{y}_{0:t-1}\right)}{\sum_j p\left(a_i = j\,|y_i, \mathbf{y}_{0:t-1}\right)}$, with the property $\pi_{ij} > 0$ and $\sum_{j=0}^{N} \pi_{ij} = 1$.
Each row in the previous matrix contains the belief for each measurement to be an outlier or the projection of each expected marker.

In the following some definitions are introduced.

Definition 1 (*Strictly Dominant Feasible Association Matrix*) The feasible association matrix $\mathcal{F}_{M_t} = \left[\,\pi_{ij}\,\right]$ is

strictly dominant if for each $i = 1, \ldots, M_t$ one j^* exists such that

$$\pi_{ij^*} > \sum_{j \neq j^*} \pi_{ij}. \tag{13}$$

The foregoing condition defines a feasible association matrix for which every measurement is univocally assigned (with a probability of more than 50%) to an outlier or to a marker.

Definition 2 (*Non-degenerate Feasible Association Matrix*) The Feasible Association Matrix $\mathcal{F}_{M_t} = \left[\pi_{ij} \right]$ is non-degenerate if it is strictly dominant and

$$\nexists j^* \neq 0 \mid \pi_{hj^*} > \sum_{j \neq j^*} \pi_{hj}, \ \pi_{ij^*} > \sum_{j \neq j^*} \pi_{ij}, \ \forall h \neq i. \tag{14}$$

The condition of non-degenerateness states that, while it is expected that more measurements can be classified as outliers ($j^* = 0$), two (or more) different measurements cannot be assigned to the same marker.

These two definitions are useful since when the property of non-degenerateness holds, the Kalman Filter correction can be performed by employing the (estimated) visible markers and their associated measurements and factoring out the measurements classified as outliers. However, ambiguities could raise when the feasible association matrix degenerates, i.e. when condition (14) is not addressed.

Experimental tests revealed that condition (13) usually holds[1], but the case of multiple association is very common and countermeasures need to be taken. In particular, it is possible that two (or more) different measurements can be assigned to the same marker. This may occur, for example, when two marker measurements are very close each other or when the same marker is split into two (or more) different measurements due, for instance, to illumination artifacts. In the degenerate situations of multiple associations, a fast and easy a posteriori algorithm is proposed in [22]. In order to disambiguate the association it has been proposed to simply take the maximum among all the π_{hj^*} and associate the marker j^* to the measurement whose probability π_{hj^*} has the maximum value. For all the remaining $h \in \mathcal{H}$ the association is forced to an outlier: $\pi_{h0} = 1$ and $\pi_{hj} = 0, \forall j = 1, \ldots, N$.

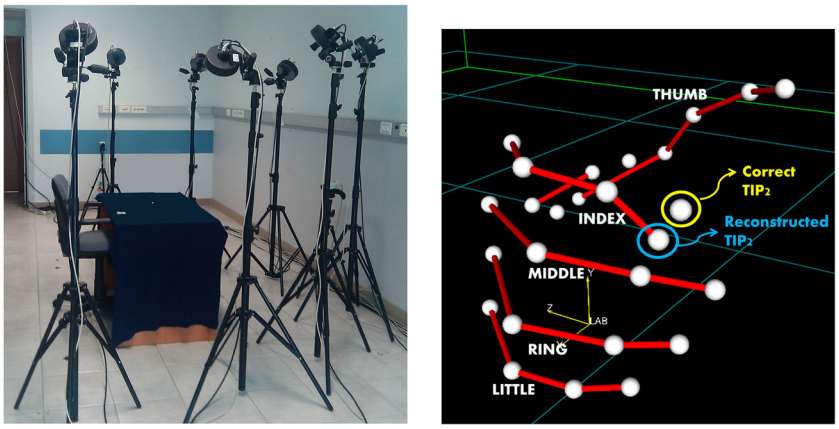
[1]Otherwise it should be hopefully possible to extract the subset of strictly dominant rows from the matrix and work with them.

3 Experimental results

3.1 Experimental setup

One volunteer subject has been involved in the experiments. It is worth observing that, for the aim of this study (consisting of assessing the accuracy and robustness of the proposed approach), the analysis of data coming from only one subject performing different grasping actions is not limiting. The subject has been asked to perform different movements with his right hand. Twenty-five reflective semispherical markers of $6mm$ diameter have been located on the hand, as shown in Fig. 1. The subject grasped three different objects with a tripod, a pinch, and a palmar grasp. He was seated in front of a table where the objects were placed one after the other. In the hand starting configuration the four fingers were fully extended and the thumb was adducted. The marker positions have been acquired with the optoelectronic system BTS SMART-D Motion Capture System (Fig. 2a), a 7-camera motion analysis system with an acquisition rate of 60 Hz. The accuracy is less than $0.1mm$ over a $2x2m$ area. The BTS Smart Analyzer software package has been used to reconstruct the marker Cartesian positions and to build a link model of the hand. The obtained marker positions constitute our 3D Cartesian coordinates ground truth. A comparison with respect to a ground truth is not possible for the angle reconstruction with the adopted instrumentation (the BTS can be used as a ground truth for the Cartesian positions but not for the joint angles which should be computed offline). In fact, the processing software provided by the optoelectronic camera manufacturers does not directly provide joint angles values, but they should be computed offline after defining, for each image frame, the joint reference frames (Fig. 3). Therefore, a synchronized acquisition with the BTS and the angular sensors embedded in a robotic hand (i.e., the DLR-HIT Hand II) has been performed. The DLR-HIT-Hand II is a dexterous robotic hand with five identical fingers and an independent palm. Each finger has four DOFs (MCP A/A and MCP, PIP, DIP F/E), of which three DOFs are actuated and one is passive. PIP and DIP joints are 1 : 1 coupled. Each finger has three Hall-effect sensors for measuring joint positions. The angular values measured by the sensors embedded in the robotic hand has been considered as the angular ground truth.

Dealing with occlusions and disappearing/reappearing markers is crucial during this phase. Standard tracking schemes sometimes may fail (Fig. 2b), thus making the labeling procedure tricky and requiring a direct intervention of the user. This makes the procedure time-consuming and the final reconstruction more error-prone. Therefore, at the beginning of the acquisition, a hand kinematic model is



(a) (b)

Fig. 2 **a** Acquisition platform composed of the BTS motion analysis system. **b** A possible association failure by using standard tracking schemes. It is evident the wrong assignment of new label (in *blue*) to the marker on index finger (in *yellow*). In this case, the user has to correct manually the software error

created and used in the estimation procedure. It guarantees that (i) differently from [11], not realistic hand configurations are avoided; (ii) the generality of the approach is increased (without resorting to ad hoc databases, as done in [10]) by developing personalized kinematic models for each subject; (iii) the estimation errors in case of occlusions or disappearing/reappearing markers is reduced by constraining the markers among each other.

The marker positions have been recorded in the starting position and during the whole trial until object grasping.

In order to measure the repeatibility of the markers positioning, the markers have been applied, removed, and repositioned five times on the subject hand and the data have been acquired with the hand in the starting configuration. During the five repositioning, the markers forming the hand reference frame have not been repositioned, in order to have a fixed reference frame with respect to obtain the marker 3D Cartesian coordinates. The difference between the 3D Cartesian coordinates of each marker in the 5 acquisitions has been evaluated and the maximum error (i.e. $1.32mm \pm 0.25mm$) has been obtained for the TMb marker.

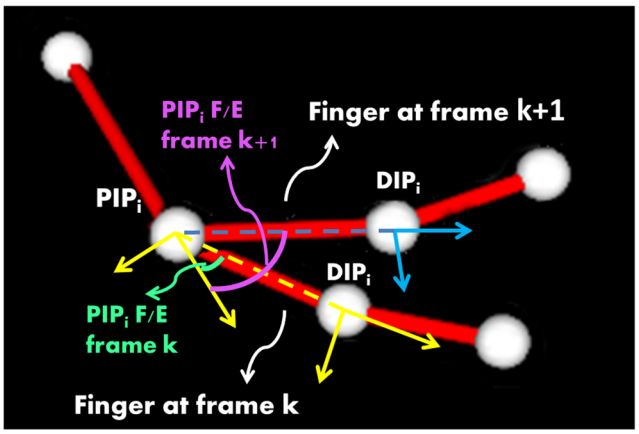


Fig. 3 Finger position in two different frames. The DIP joint reference frame is outlined in *yellow* at frame k, whereas it is in *blue* at frame $k+1$. The proposed protocol determines joint reference frames at each video frame. The relative position of two consecutive joint reference frames gives joint angle values. PIP F/E angle is in *green* at frame k, in *violet* at frame $k+1$

3.2 Initialization phase

An initialization phase needs to be performed to build the hand kinematic model and estimate the initial relative transformation between camera and hand reference frame. At the beginning of the acquisition (i.e. in the first frame), a Non Linear Least Squares optimization approach has been applied to the measurements of the hand markers. It is paramount that during this phase all the markers are visible and correctly associated to the corresponding measurements.

For the initialization of position and orientation of the hand frame, let $T_{m_i b}$ be the position of the i-th marker on the dorsum and $\widetilde{y}_{m_i c}$ be the corresponding 3D measurement in the camera frame. The available measurements are mapped into the estimated positions of the markers in the camera frame through the following relationship

$$y_{m_i c} = g_0 T_{m_i b} \tag{15}$$

where $g_0 \in SE(3)$ represents the initial pose of the hand with respect to the camera. The 2-norm cost function

$$\widehat{g}_0 = \min_{g_0} \sum_i \| \widetilde{y}_{m_i c} - g_0 T_{m_i b} \|^2. \tag{16}$$

is minimized in order to find the optimal estimation of the foregoing transformation. As regards the initialization of the finger joint angles, the transformation that maps this position onto the corresponding measurement in the camera frame can be written as

$$y_{m_i c} = \widehat{g}_0 T_{m_i b} \left(\Theta_0^j \right), \tag{17}$$

where $T_{m_i b}\left(\Theta_0^j\right)$ is the position of the i-th marker on the j-th finger with respect to the frame on $B1$ and depends on the configuration of the finger, i.e. on the values Θ_0^j of the joint angles. The locally optimal estimation of the initial configuration of the finger, Θ_0^j, is found by minimizing the 2-norm cost function

$$\widehat{\Theta_0^j} = \min_{\Theta_0^j} \sum_i \|\widetilde{y}_{m_i c} - \widehat{g}_0 T_{m_i b}(\Theta_0^j)\|^2. \tag{18}$$

These initial motion and pose parameters are used to initialize the state of the Unscented Kalman Filter and to compute the hand kinematic model.

3.3 Validation of the stochastic algorithm for the estimation of marker 3D Cartesian coordinates

Unknown associations have been solved with the described probabilistic association method. It is necessary to outline that, for the stochastic algorithm, all the measures provided by the optoelectronic cameras have been used without applying the processing software provided by the optoelectronic camera manufacturers. Furthermore, the stochastic approach is based on a fixed hand reference frame centered on the marker $B1$ (as explained in Section 2.1), defined at the first frame during the calibration phase; the Cartesian coordinates of all the hand joints are expressed with respect to this fixed hand reference frame.

In order to validate the approach, algorithm outcomes (in terms of joint Cartesian coordinates) have been compared with data acquired with the optoelectronic system and reconstructed with the BTS Smart Analyzer (i.e., our ground truth). The mean and standard deviation of the error between joint Cartesian coordinates obtained with the two methods, computed in all the frames, are shown in Fig. 4. For the sake of brevity, the figure shows the performance analysis, in terms of error between the finger joint Cartesian coordinates only during the tripod grasp, but similar results have been also obtained for the other grasping configurations. The red dots represent the mean value and the blue lines the standard deviation obtained for each Cartesian coordinate of each finger joint. As it can be seen, the error mean value is always less than $4mm$ (i.e., it is maximum $3mm \pm 0.9mm$) demonstrating the good performance, in terms of Cartesian position reconstruction accuracy, of the proposed estimation approach.

During the acquisitions, some occlusions and disappearing/reappearing markers phenomena happened. The results shown in Fig. 4 confirm the reliability of the proposed method also in presence of these phenomena.

A representative frame of the comparative analysis between the methods in terms of joint Cartesian coordinates is shown in Fig. 5. White dots are the joint Cartesian coordinates given by the optoelectronic system and red squares are the joint Cartesian coordinates resulting from the stochastic approach.

The algorithmic complexity of the stochastic filter with the robust tracking scheme has also been evaluated by measuring the execution time of the algorithm during the experiments. The average value of the execution time of the algorithm running in Matlab is $50ms$ (± 6). This reveals

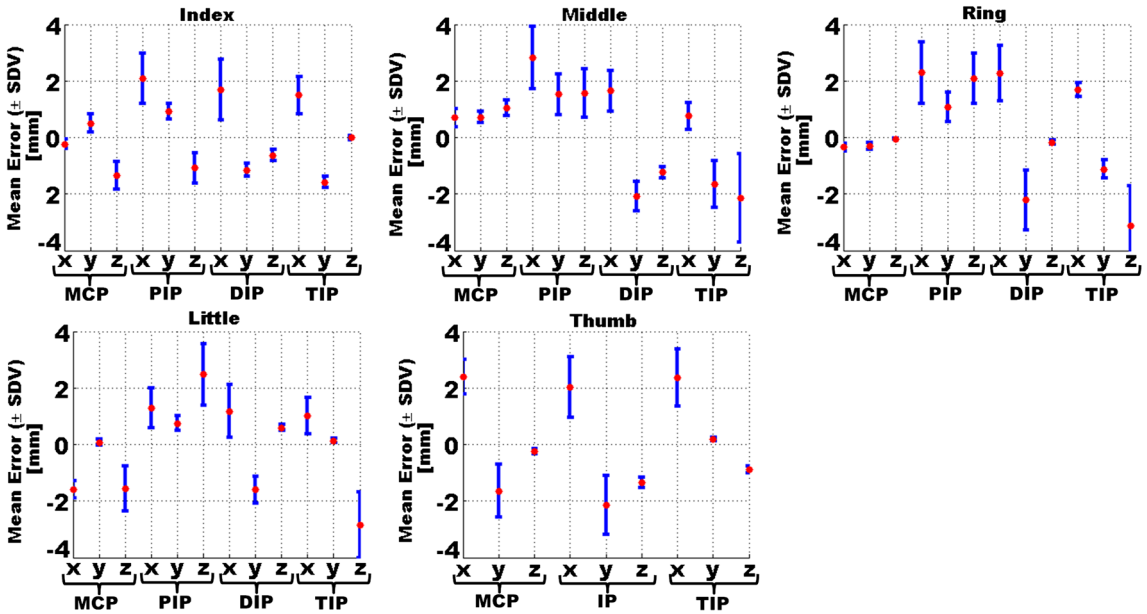


Fig. 4 Mean value (± standard deviation) of the error between joint Cartesian coordinates obtained with the stochastic algorithm and with the BTS Smart Analyzer, calculated in all the frames in the case of a tripod grasp.

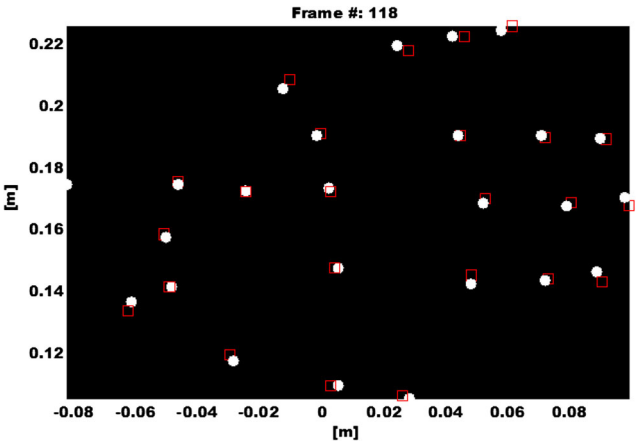


Fig. 5 Marker positions given by the BTS Smart Analyzer (outlined in *white*) and by the stochastic algorithm (in *red*).

that, subject to some improvement, an implementation of the algorithm in real-time is feasible.

3.4 Validation of the stochastic algorithm for the estimation of hand joint angles

In order to validate the joint angle reconstruction, 21 retroreflective markers have been located on the robotic hand joints adopting the same protocol applied to the human hand (Fig. 6).

The index finger has been moved from an initial angle configuration equal to: $[MCP_{A/A} = 0rad;\ MCP_{F/E} = 0rad;\ PIP_{F/E} = 0rad]$ to a final angle configuration equal to: $[MCP_{A/A} = 0rad;\ MCP_{F/E} = 0.61rad;\ PIP_{F/E} = 0.78rad]$. A third degree polynomial function and a PD torque control in the joint space have been used to control the hand index finger closing until the final desired configuration. The joint angles measured by the sensors embedded in the robotic hand represent our ground truth.

In order to evaluate the performance of the proposed stochastic algorithm, a comparison with the angular ground truth has been performed. In particular, the error between the angular values obtained by means of the angular sensors embedded in the robotic hand and the one estimated by the stochastic algorithm has been evaluated. The error behavior is shown in Fig. 7. The 1:1 coupling between the DIP and PIP joints has been modeled also in the kinematic model for the stochastic algorithm. Therefore, the error on the DIP joint is equal to the error on the PIP joint.

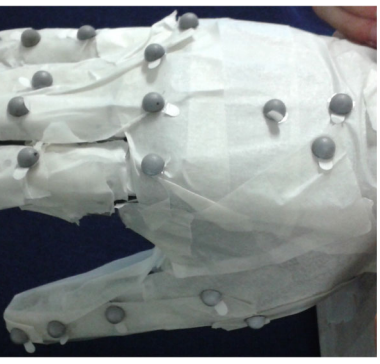


Fig. 6 Adopted protocol for marker positioning on the robotic hand. The hand has been covered with paper tape to avoid interferences between the camera infrared rays and the metallic cover of the hand.

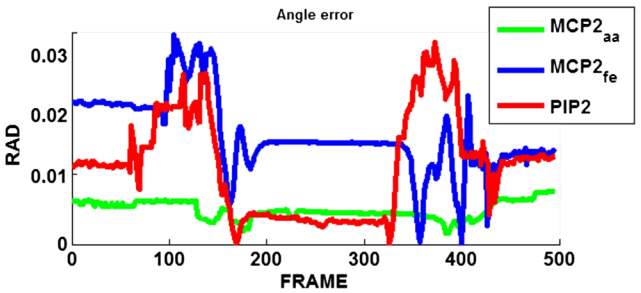


Fig. 7 Error between the angular values obtained by means of the angular sensors embedded in the robotic hand and the one estimated by the stochastic algorithm. The error behavior is shown for the $MCP_{A/A}, MCP_{F/E}, PIP_{F/E}$

It is possible to note that the error is always less than $0.035rad \pm 0.006rad$ over a range of motion of $0rad - \pi rad$ and is better than the results obtained in the literature [23].

4 Discussion

It is worthwhile mentioning that in a previous paper [20], a preliminary version of the stochastic approach applied to a RGB-D camera was presented. In the present paper, an evolution of the approach and its application to a multi-camera optoelectronic system are illustrated. A more rigorous formalization of the method has been developed and the problem of associating different measurements to the same marker has been solved. Furthermore, the system performance has been thoroughly tested by means of a comparison with ground truths of angular and 3D coordinate measures.

The results clearly show that the pose estimation algorithms lead to low errors between the estimated values and the ground truth both as regards joint Cartesian coordinates (maximum $3mm \pm 0.9mm$, which is a typical error of the estimation algorithms proposed in the literature) and joint angle values (less than $0.035rad \pm 0.006rad$). In particular, from Fig. 4, it is evident that the largest error is on the z-component of the TIPs. This is probably due to the misalignment of the markers with respect to the corresponding joint (whereas in this work they are supposed coincident) and to finger anatomy, which is different from the modeled one: fingers actually are not cylinders, but truncated cones slightly curved along the z-axis. Furthermore, there is also a

small error propagation along the kinematic chain. Thus, the error between the estimated joint Cartesian coordinates and the measurements are related to the errors due to the pose estimation and to the forward kinematic approximations. The mean error on the TIP Cartesian coordinates obtained with the proposed stochastic method is comparable or even lower than the error obtained by other approaches proposed in the literature, where data provided by an optoelectronic camera system have been used as ground truth. In [24], the error between the TIP Cartesian coordinates estimated by their inertial sensory system and those measured by means of the optoelectronic system reaches $10mm$ on the z-direction. In [7], the maximum error between joint Cartesian coordinates predicted by their model and the coordinates measured by the optoelectronic system is about $3.5mm$ on the TIP.

The comparison with the angular ground truth let us to state that this method is also applicable to activities requiring an accurate angle estimation, such as patient performance evaluation. The stochastic filter is fast, probabilistically robust and optimal. The results show robustness and real-time applicability of the proposed approach. Moreover, it is very easy to be implemented and initialized, and requires a minimum effort by the user, as several experiments have shown. This last consideration makes the stochastic approach feasible for biomedical applications, such as neurorehabilitation: e.g. the therapist motion can be reconstructed to guide a rehabilitation robot or the motor tasks executed by the patient can be used for moving a patient avatar in a virtual environment during a rehabilitation session.

5 Conclusions

An automatic, robust and simple method for hand pose estimation using a multi-camera system and a model-based stochastic algorithm has been presented. The marker-based approach relies on a UKF and an accurate hand kinematic model in order to improve the algorithm robustness with respect to outliers and possible occlusions. The approach has been validated by analyzing data acquired with an optoelectronic system during three different types of grasp. It has been compared with ground truths for angular and 3D coordinate measures. The comparative analysis shows a level of accuracy in the reconstruction of the joint Cartesian coordinates comparable or better with respect to other methods proposed in the literature (less than $4mm$). The results obtained for the joint angles reconstruction are realistic and within errors comparable to those reported in the literature (less than $0.035rad$). The advantages of using the proposed model-based stochastic algorithm with respect to other optoelectronic methods for hand joint motion reconstruction are evident in terms of simplicity of implementation, initialization, effort by the user and possibility to use the approach in real-time, making the system suitable for several biomedical applications. Namely, the performance achieved with our stochastic approach offers the chance to use the method in the field of neurorehabilitation combined with virtual reality, where patients have to perform predefined hand movements and visualize their hand avatar in real-time, and with instrumented objects for acquiring information about the interaction forces between the human hand and the grasped objects [25, 26]. Future activities will be mainly focused on the application of the approach to the hand motion analysis during rehabilitation sessions.

Acknowledgments This work was supported partly by the Italian Institute for Labour Accidents (INAIL) with PPR 2 project (CUP: E58C13000990001) and partly by the European Project H2020/AIDE: Multimodal and Natural computer interaction Adaptive Multimodal Interfaces to Assist Disabled People in Daily Activities (CUP J42I15000030006).

Appendix

The Denavit-Hartenberg (DH) parameters for the index finger and for the thumb are shown in Tables 1 and 2, respectively. The other long fingers have the same DH parameters of the index. DH parameters are evaluated in such a way as to obtain a generic algorithm valid for different hand sizes. Therefore, the algorithm envisages an initial calibration phase, where the 3D Cartesian coordinates of the markers center are detected manually in the first image acquired by the camera and the link lengths are measured, by means of the 3-dimensional information provided by the vision system.

In the Tables 1 and 2, L^{index} and L^{thumb} represent the link lengths of the index finger and of the thumb, respectively.

Once the DH parameters have been computed, the rotation matrices can be extracted. Given the symbolic form of a generic rotation matrix

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \tag{19}$$

Table 1 Denavit-Hartenberg parameters of the index finger

Joint #	a_i	α_i	d_i	θ_i
1	0	$-\pi/2$	0	q_{flex}^{MCP}
2	L_P^{index}	$\pi/2$	0	q_{abd}^{MCP}
3	L_M^{index}	0	0	q_{flex}^{PIP}
4	L_D^{index}	0	0	q_{flex}^{DIP}

Table 2 Denavit-Hartenberg parameters of the thumb

Joint #	a_i	α_i	d_i	θ_i
1	0	$-\pi/2$	0	q_{flex}^{TM}
2	L_P^{thumb}	$\pi/2$	0	q_{abd}^{TM}
3	0	$-\pi/2$	0	q_{flex}^{MCP}
4	L_M^{thumb}	$\pi/2$	0	q_{abd}^{MCP}
5	L_D^{thumb}	0	0	q_{flex}^{IP}

the corresponding Euler angles in configuration ZYX, under the assumption that $r_{13} \neq 0$ and $r_{23} \neq 0$, are

$$\begin{aligned} \phi &= atan2(r_{23}, r_{13}) \\ \theta &= atan2(\sqrt{r_{13}^2 + r_{23}^2}, r_{33}) \\ \psi &= atan2(r_{32}, -r_{31}) \end{aligned} \tag{20}$$

where $atan2(x, y)$ is the arctangent of two arguments, the choice of the positive sign for the term $\sqrt{r_{13}^2 + r_{23}^2}$ limits the range of the feasible values of θ to $(0, \pi)$. If θ is chosen in the range $(-\pi, 0)$, Eq. 20 becames

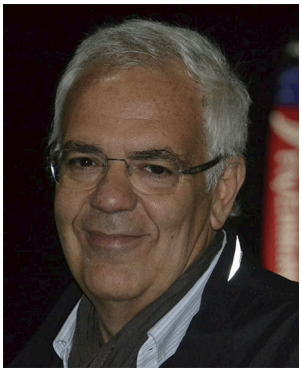
$$\begin{aligned} \phi &= atan2(-r_{23}, -r_{13}) \\ \theta &= atan2(-\sqrt{r_{13}^2 + r_{23}^2}, r_{33}) \\ \psi &= atan2(-r_{32}, r_{31}) \end{aligned} \tag{21}$$

References

1. Provenzale A, Cordella F, Zollo L, Davalli A, Sacchetti R, Guglielmelli E (2014) A grasp synthesis algorithm based on postural synergies for an anthropomorphic arm-hand robotic system. Proc IEEE RAS EMBS Int Conf Biomed Robot Biomechatron
2. Foxlin E (2007) Motion tracking requirements and technologies. In: Handbook of virtual environments: design, implementation, and applications, Lawrence Erlbaum Associates, vol 35, pp 1989–2002
3. Bianchi M, Salaris P, Bicchi A (2013) Synergy-based hand pose sensing: optimal glove design. Int J Robot Res 32:396–406. doi:10.1177/0278364912474079
4. Oikonomidis I (2012) Tracking the articulated motion of two strongly interacting hands. Proc CVPR IEEE, Argyros AA
5. Sridhar S, Mueller F, Oulasvirta A, Theobalt C (2015) Fast and robust hand tracking using detection-guided optimization. In: Conference on computer vision and pattern recognition, pp 3213–3221
6. Wheatland N, Wang Y, Song H, Neff M, Zordan V, Jorg S (2015) State of the art in hand and finger modeling and animation. Eurographics
7. Cerveri P, De Momi E, Lopomo N, Baud-Bovy G, Barros RML, Ferrigno G (2007) Finger kinematic modeling and real-time hand motion estimation. Ann Biomed Eng 35:11. doi:10.1007/s10439-007-9364-0
8. Meyer J, Kuderer M, Muller J, Burgard W (2014) Online marker labeling for fully automatic skeleton tracking in optical motion capture. IEEE Int Conf Robot Autom
9. Zordan VB, Van Der Horst NC (2003) Mapping optical motion capture data to skeletal motion using a physical model. In: ACM SIGGRAPH/Eurographics symposium on computer animation
10. Maycock J, Rohlig T, Schroder M, Botsch M, Ritter H (2015) Fully automatic optical motion tracking using an inverse kinematics approach. In: IEEE-RAS international conference on humanoid robots
11. Aristidou A, Lasenby J (2013) Real-time marker prediction and cor estimation in optical motion capture. Vis Comput 29:7–26. doi:10.1007/s00371-011-0671-y
12. Kandepu R, Foss B, Imsland L (2008) Applying the unscented Kalman filter for nonlinear state estimation. J Process Control 18:753–768. doi:10.1016/j.jprocont.2007.11.004
13. Liu H, Wu K, Meusel P, Seitz N, Hirzinger G, Jin MH, Liu YW, Fan SW, Lan T, Chen ZP (2008) Multisensory five-finger dexterous hand: the DLR/HIT Hand II. IEEE Int C Int Robot
14. Bullock IM, Borras J, Dollar AM (2012) Assessing assumption in kinematic hand models: a review. Proc IEEE RAS EMBS Int Conf Biomed Robot Biomechatron
15. Cobos S, Ferre M, Uran S, Ortego J, Pena C (2008) Efficient human hand kinematics for manipulation tasks. IEEE Int C Int Robot
16. Chang LY, Pollard NS (2008) Method for determining kinematic parameters of the in vivo thumb carpometacarpal joint. IEEE Trans Biomed Eng 55:1897–1907. doi:10.1109/TBME.2008.919854
17. Cordella F, Zollo L, Salerno A, Accoto D, Guglielmelli E, Siciliano B (2014) Human hand motion analysis and synthesis of optimal power grasps for a robotic hand. Int J Adv Rob Syst 11:1–13. doi:10.5772/57554
18. Siciliano B, Sciavicco L, Villani L, Oriolo G (2009) Robotics — modelling, planning and control. Springer, London. ISBN 978-1-84628-641-4
19. Wan EA, Van der Merwe R (2000) The unscented Kalman filter for nonlinear estimation. In: Symposium on adaptive systems for signal processing, communications, and control
20. Cordella F, Di Corato F, Zollo L, Siciliano B (2013) New trends in image analysis and processing ICIAP 2013, Lecture Notes in Computer Science. In: Petrosino A, Maddalena L, Pala P (eds) A robust hand pose estimation algorithm for hand rehabilitation. Springer Verlag, Berlin, pp 1–10
21. Sarkka S, Vehtari A, Lampinen J (2004) Raoblackwellized Monte Carlo data association for multiple target tracking. In: Proceedings of the 7th international conference on information fusion, vol 1, pp 583–590
22. Di Corato F (2013) A unified framework for constrained visual-inertial navigation with guaranteed convergence. PhD Dissertation, University of Pisa
23. Liang H, Yuan J, Thalmann D, Zhang Z (2013) Model-based hand pose estimation via spatial-temporal hand parsing and 3D fingertip localization. Vis Comput 29:837–848
24. Kortier HG, Sluiter VI, Roetenberg D, Veltink PH (2014) Assessment of hand kinematic using inertial and magnetic sensors. J Neuroeng Rehabil 11:70. doi:10.1186/1743-0003-11-70
25. Cordella F, Taffoni F, Raiano L, Carpino G, Pantoni M, Zollo L, Schena E, Guglielmelli E, Formica D (2016) Design and development of a sensorized cylindrical object for grasping assessment. Conf Proc IEEE Eng Med Biol Soc
26. Romeo RA, Cordella F, Zollo L, Formica D, Saccomandi P, Schena E, Carpino G, Davalli A, Sacchetti R, Guglielmelli E (2015) Development and preliminary testing of an instrumented object for force analysis during grasping. Conf Proc IEEE Eng Med Biol Soc

Francesca Cordella has a Ph.D. in Computer and Automation Engineering and is the Research Assistant of biomedical robotics at Università Campus Bio-Medico of Rome. Her research interests include biomechanics and robotics.

Francesco Di Corato received the M.Sc. degree in Automation & Robotics Engineering and the Ph.D. in Automation, Robotics & Bioengineering from the University of Pisa. He is a Sr. Automotive Research Engineer at Vislab.



Bruno Siciliano is a Full Professor of Robotics. He is the director of the ICAROS and responsible of the PRISMA Lab. He has delivered more than 120 keynotes and has published more than 300 papers and 7 books.

Loredana Zollo is an Associate Professor of Bio-Engineering. She has authored more than 90 peer-reviewed publications which appeared in journals, books, and conference proceedings in the fields of biomedical and neuro-robotics.