



A Novel Approach for Cooperative Motion Capture (COMOCAP)

Gregory Welch^{1,3} , Tianren Wang², Gary Bishop³, and Gerd Bruder¹ 

¹University of Central Florida, USA

²RoboteX, Inc., USA

³University of North Carolina at Chapel Hill, USA

Abstract

Conventional motion capture (MOCAP) systems, e.g., optical systems, typically perform well for one person, but less so for multiple people in close proximity. Measurement quality can decline with distance, and even drop out as source/sensor components are occluded by nearby people. Furthermore, conventional optical MOCAP systems estimate body posture using a global estimation approach employing cameras that are fixed in the environment, typically at a distance such that one person or object can easily occlude another, and the relative error between tracked objects in the scene can increase as they move farther from the cameras and/or closer to each other. Body-relative tracking approaches use body-worn sensors and/or sources to track limbs with respect to the head or torso, for example, taking advantage of the proximity of limbs to the body. We present a novel approach to MOCAP that combines and extends conventional global and body-relative approaches by distributing both sensing and active signaling over each person's body to facilitate body-relative (intra-user) MOCAP for one person and body-body (inter-user) MOCAP for multiple people, in an approach we call cooperative motion capture (COMOCAP). We support the validity of the approach with simulation results from a system comprised of acoustic transceivers (receiver-transmitter units) that provide inter-transceiver range measurements. Optical, magnetic, and other types of transceivers could also be used. Our simulations demonstrate the advantages of this approach to effectively improve accuracy and robustness to occlusions in situations of close proximity between multiple persons.

CCS Concepts

• **Human-centered computing** → *Mixed / augmented reality; Virtual reality; Graphics input devices*; • **Computing methodologies** → *Motion capture; Graphics input devices; Mixed / augmented reality; Virtual reality; Motion capture*;

1. Introduction

There are many techniques and systems designed for human motion capture (MOCAP), which are widely used for animating film and video game characters, assessing human movement in healthcare situations, sports analysis, and a wide range of training activities [MHK06, WF02]. Systems exist that employ inertial sensors, e.g., [DOKA13, Not], but the most popular are optical or magnetic systems that employ environment-mounted cameras or magnetic sources that “look” or transmit inward toward human subjects who are wearing passive optical reflectors [ART17, Opt17a, VIC], active light sources [Opt17b, Pha17], or magnetic sensors [Pol17]. However, existing systems have difficulty with multiple people being captured in the same space. When people are tracked with respect to components that are fixed in the environment, the relative error between tracked people will increase as they move farther from the environment-mounted sources/sensors and/or closer to each other. Furthermore, as people or objects move closer to each other they can completely block signals to/from the environment-mounted components. Unfortunately these occlusion and distance-related er-

rors tend to increase under exactly the circumstances when the relative accuracy matters the most—when captured people are interacting close together.

In this paper, we introduce the notion of *cooperative motion capture* (COMOCAP) for circumstances where proximal interactions between multiple users are expected. As illustrated in Figure 1, the basic idea is to replace or supplement conventional environment-mounted and body-worn components with environment-mounted and body-worn *transceivers* (transmitter-receiver units) used to cooperatively measure inter-transceiver geometric relationships, and to use those measurements to continuously jointly estimate the evolving body postures with respect to the fixed environment (global), *within* users (intra-user), and *between* users (inter-user).

The cooperative nature of the approach requires transceivers that can signal each other in both directions. This can be achieved for example with optical, magnetic, or acoustic sources and sensors. Compared to global posture measurements obtained from fixed sensors, e.g., room/tripod-mounted cameras, cooperative intra-body or inter-body measurements between body-mounted

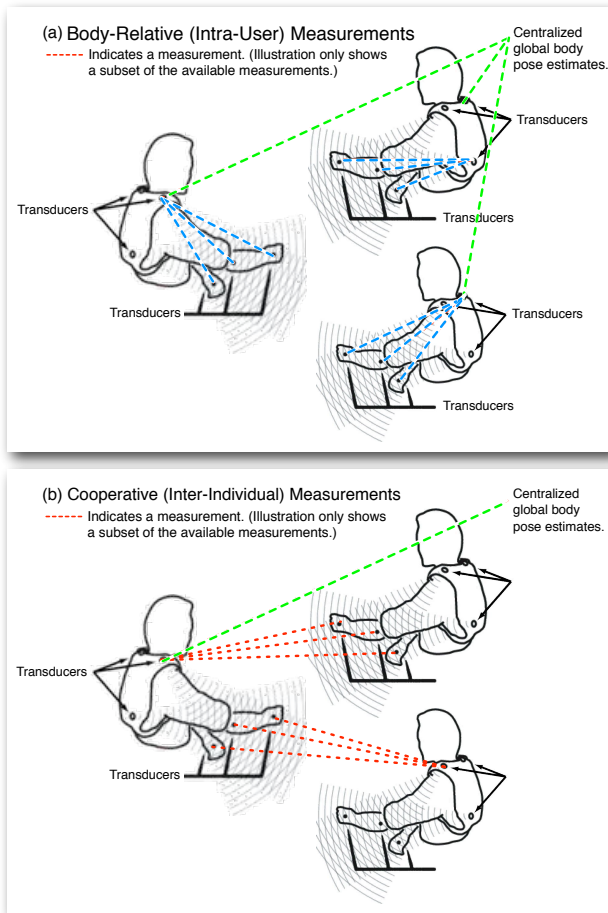


Figure 1: (a) A depiction of body-relative signaling in a multi-person setup. Each person wears components that can send/receive intra-user signals (blue dashed) between a reference body point, typically the head or torso, and the moving limbs. Each person's body is then tracked with respect to the environment via a centralized global system (green dashed). (b) A depiction of our proposed COMOCAP approach where transceivers (transmitter-receiver units) worn on each person send/receive signals (red dashed) to/from the transceivers of other people in a peer-to-peer fashion.

transceivers can increase robustness in the face of occlusions or other measurement failures that would occur with passive or independently-functioning active markers in such circumstances. The cooperative measurements can also stabilize or reduce the relative error between body parts, especially in proximal conditions.

We evaluated this theoretical model with respect to a simulation of an acoustic system employing conventional and COMOCAP approaches. Our simulation results support the validity of the COMOCAP approach, indicating increased accuracy when people are close to each other. Our results also support the notion that the COMOCAP approach can improve the robustness to sensor occlusion compared to conventional approaches.

2. Related Work

Motion tracking technologies aimed at sensing the pose of a human body can be categorized in multiple ways, such as *body-mounted* or

environment-mounted sensors or sources, *marker-based* or *marker-less* systems, and *physical contact-based* or *contact-less* approaches [MHK06, WF02]. Moreover, motion tracking approaches can be classified based on the medium that is used for measurements, such as mechanical [Ani17], inertial [DOKA13, Neu17], acoustic [VAV*07], magnetic [Pol17], optical [Atr17, Opt17a], and radio frequency [KPP*13]. In an effort to remedy shortcomings in some of these mediums, hybrid approaches have been proposed based on separate position/orientation sensors or multiple overlapping data sources with sensor fusion [HKS*15, HSGS06, ZP17]. While our general COMOCAP approach is not limited, we limit the scope of this paper to a single-medium system, and give an outlook on potential hybrid approaches in our discussion of future work.

In this paper we distinguish *environment-reference* from *body-reference* approaches. Environment-reference tracking approaches measure angles or distances from fixed sensors in the environment to the user's body and then interpret the signals to estimate the body pose and movement. The most prominent examples of motion capture systems that adopt environment-reference measurements are based on optical sensors (e.g., 4D light-field sensors [JSG15], 2D CCD or CMOS [DU02, DB03, Osh06], or 1D line cameras [Atr17]). Most optical motion capture systems fixed cameras mounted in the environment to observe retroreflective or active markers placed on the human body [ART17, Opt17a, Opt17b, Pha17, VIC]. Marker-less motion capture systems typically use active depth (e.g., Kinect) or image-based silhouettes of human bodies [Lok02, RKS*05, Org17]. These approaches suffer from occlusion in general and with multiple users in particular.

Much previous research has been dedicated to determining an optimal camera placement to provide an unobstructed path between cameras and users [RK17]. The situation can be modeled as a visibility or *Art Gallery* problem if the occluders are static [FCOL99], but these solutions fail for dynamic occluders, e.g., due to one's own body movements or the movement of another human's body in the tracked space. This is true even for acoustic tracking systems such as *Whisper* [Val02], which uses a wide bandwidth signal to take advantage of a low frequency sound's ability to diffract around occluders. Tracking multiple users simultaneously moving close to each other in a typical 5m×5m room introduces intractable infrastructure challenges for such environment-reference optical motion capture systems. For instance, following the VICON Full-Body Animation configurator [VIC], motion capture for one user requires 10 cameras (5m×5m), five users 24 cameras (9m×9m), and ten users 36 cameras (15m×15m).

Body-reference tracking approaches measure angles or distances between sensors fixed on the human body. For instance, inertial units mounted near bones or joints of a human body can be used to measure relative orientations and/or positions along the kinematic chain of the body [DOKA13]. Recent consumer body tracking systems using this approach include *Notch*, *Perception Neuron*, and *PrioVR*. Acoustic sensors can help correct for drift errors in inertial measurements [VAV*07]. Since there is no need for external references, these approaches offer useful solutions for motion capture in everyday environments and can be set up outdoors. To reduce drift, body-reference tracking systems sometimes integrate analyti-

cal priors based on kinematic constraints, or seek to integrate additional environment-reference sensing technologies (see Figure 1a).

Related work has been conducted by Johnson et al., who designed a cooperative multi-projector pose estimation framework capable of calibrating the 3D pose of every projector-camera (Pro-Cam) unit within an environment via image patterns not only projected by its own, but also projected by other units [JWF*09]. Our approach is related to 1991 work by Rao and Durrant-Whyte [RDW91], who describe how to decentralize tasks among several sensor units and then assimilate each unit's result so that every unit arrives at a global common consensus, which enables systems that are robust and do not need a central processing unit.

3. Method

The basic idea of our COMOCAP approach is to replace or supplement conventional environment-mounted and body-worn components with environment-mounted and body-worn *transceivers* used to cooperatively measure inter-transceiver geometric relationships, with every environment/user transceiver would transmitting/receiving to/from all other environment/user transceivers nearby. One could also combine user-worn transceivers with a conventional environment-mounted MOCAP system.

There are several advantages to using the cooperative approach, in particular when there are multiple nearby or interacting users:

1. the additional inter-individual cooperative measurements provide *increased* pose/posture information over current approaches, which can increase accuracy and robustness;
2. the inter-individual cooperative measurements can provide information when global sensors are otherwise occluded—self occluded or occluded by others; and
3. the *direct* nature of inter-individual cooperative measurements can *reduce* the otherwise typically increasing relative error as users move closer to each other.

For our COMOCAP implementation we employ an extended Kalman filter [Kal60, WB95] with a *position-velocity* (PV) model as taught in [ABW01, BH96]. The EKF is attractive because it can weigh noisy measurements against a model for the expected motion and fuse them together to arrive at theoretically optimal pose estimates, and it maintains an estimate of the state error covariance which provides an ongoing indication of the quality of the actual and expected pose estimation—particularly valuable for measurement selection from a set of alternatives as we discuss below.

While the notion of global estimates of optical marker positions in 3D is relatively straight forward, the one-to-many nature of inter-individual transceiver measurements can make explanations confusing. In particular, consider that a single acoustic transceiver can transmit a sound that can be received by multiple other acoustic transceivers, resulting in multiple simultaneous range measurements. Furthermore in the case of acoustic transceivers these one-to-many measurements are *relative* range measurements between transceivers that simply constrain each transceiver to be somewhere on a 3D sphere around itself or the other transceiver.

At every estimate cycle it is useful to identify the set of transducers associated with one user as the *target* transducers U_T , while

referring to all others as the *reference* transducers U_{R_i} , where $i = 1, 2, \dots, v - 1$ for v users. We use τ to denote the number of transceivers worn by each user, and ϵ the number of transceivers in the environment (fixed known locations). For the sake of explanation we assume a constant EKF update rate of $1/\delta t$ updates per second (δt seconds between updates), and per Kalman filter conventions use the subscripts k to indicate the current time step (i.e. time t) and $k - 1$ to indicate the previous time step (i.e. time $t - \delta t$).

We implement a single EKF per user, with the user's state representing the collective state of all of the transceivers worn by that user. Specifically, if one considers a 6-dimensional state vector

$$X_i = [x_i, \dot{x}_i, y_i, \dot{y}_i, z_i, \dot{z}_i]^T, \quad (1)$$

for each transceiver, then each user has an associated collective 6τ -dimensional user-specific state vector

$$X^U = [X_1, X_2, \dots, X_\tau]^T. \quad (2)$$

Similarly each transceiver has an associated 6×6 error covariance P_i ($i = 1, 2, \dots, \tau$), which collectively form a $6\tau \times 6\tau$ error covariance matrix P^U for each user.

At each time step k of the EKF we choose a single *target* user denoted by the superscript U_T , and define the remaining $v - 1$ users as *reference* users denoted by the superscript U_{R_i} . We combine the 6τ -dimensional per-user state vectors into an $6\tau v$ -dimensional aggregate state vector

$$\mathbb{X}_k = [X_k^{U_T}, X_k^{U_{R_1}}, X_k^{U_{R_2}}, \dots, X_k^{U_{R_{v-1}}}]^T, \quad (3)$$

and we combine the per-user $6\tau \times 6\tau$ covariance matrices into an $6\tau v \times 6\tau v$ aggregate covariance matrix \mathbb{P}_k with the error *autocovariances* for the target user filter and each reference user filter are on the diagonal. We *do* maintain the error covariances between the target user filter and each reference user filter in the first row and first column. We *do not* maintain error covariances between the reference user filters—i.e. we model the reference user filters as being independent of each other.

3.1. Time Update

In the time update step of the EKF we use a time-invariant aggregate $6\tau v \times 6\tau v$ *state transition matrix* \mathbb{A} to project the aggregate state Equation (3) and aggregate error covariance \mathbb{P}_k forward from the previous time step to the current time step, to obtain *a priori* estimates of the same as indicated by the “-” superscripts:

$$\mathbb{X}_k^- = \mathbb{A}\mathbb{X}_{k-1}, \quad (4)$$

$$\mathbb{P}_k^- = \mathbb{A}\mathbb{P}_{k-1}\mathbb{A}^T + \mathbb{Q}. \quad (5)$$

where \mathbb{A} and \mathbb{Q} are formed as follows. The time-invariant aggregate state transition matrix \mathbb{A} is a $6\tau v \times 6\tau v$ block-diagonal matrix with diagonal elements $A^{U_T}, A^{U_{R_1}}, A^{U_{R_2}}, \dots$, and $A^{U_{R_{v-1}}}$ corresponding to the $6\tau v$ -dimensional aggregate state vector in Equation (3). Each of the elements A^U has to transform the states of all of the τ transceivers worn by the associated user. However because we expect all *users* to behave with similar dynamics over time, the state transitions are identical in form. In fact because we expect all *transceivers* of each user to behave with similar dynamics over time, their state transitions are also identical. The state transition matrix corresponding to a PV dynamic model corresponding to the

state in Equation (1) would be a block diagonal matrix with three identical blocks

$$A = \begin{bmatrix} 1 & \delta t \\ 0 & 1 \end{bmatrix}, \quad (6)$$

for x , y , and z . Moving up to the level of a user with state as in Equation (2), the state transition matrix would be formed as a block-diagonal series of τ copies of A in Equation (6). These are then substituted back into the ν block-diagonal elements of \mathbb{A} above and used to transition the aggregate state and covariance matrices in Equation (4) and Equation (5).

Because we expect all *users* and *transceivers* to behave with similar dynamics over time, the aggregate $6\tau\nu \times 6\tau\nu$ block-diagonal *process noise matrix* \mathbb{Q} is both time-invariant and formed from a series of identically constructed block elements Q^{U_τ} , $Q^{U_{R_1}}$, ..., and $Q^{U_{R_{\nu-1}}}$, each assembled from 6τ block-diagonal elements Q formed as

$$Q[1,1] = q \frac{(\delta t)^3}{3}, \quad (7)$$

$$Q[1,2] = Q[2,1] = q \frac{(\delta t)^2}{2}, \text{ and} \quad (8)$$

$$Q[2,2] = q\delta t. \quad (9)$$

Each such 2×2 block element formed from Equations (7)–(9) models the process noise for one dimension (x , y , or z) of the user-specific state in Equation (2). While a more complete explanation for the elements of Equation (9) can be found in [BH96], the basic idea is that the Kalman filter assumes the process is stimulated or “fed” by a normally-distributed, zero-mean, spectrally white, process noise q . We used the method from [WB01] applied to a PV process model to choose q , and build the aggregate block diagonal $6\tau\nu \times 6\tau\nu$ process noise matrix \mathbb{Q} from 3ν copies of the 2×2 block element from Equation (9)—three 2×2 blocks per transceiver, times τ transceivers, times ν users.

3.2. Measurement Update

In the measurement update step of the EKF we collect all of the individual measurements from the environment and reference user transceivers associated with the target user, and fuse them with the *a priori* aggregate state and error covariance estimates \mathbb{X}_k^- and \mathbb{P}_k^- for that user, obtaining *a posteriori* aggregate state and error covariance estimates \mathbb{X}_k and \mathbb{P}_k .

For acoustic transceivers, each measurement from one transceiver to another is a scalar distance. Depending on the number of users (ν), transceivers per user (τ), and transceivers in the environment (ϵ), the number of possible measurements (transceiver combinations) could be quite large. For the sake of completeness we describe all combinations here, however in practice we prioritize and limit the measurements as described later in the paper. Note also that for an acoustic system, a transmission (sound) emanating from one transceiver could possibly be received (“heard”) by all other transceivers on the target user and the reference users, and fixed in the environment, offering significant measurement per time efficiencies.

Like a conventional global approach one can acquire measurements between the τ target user-worn and ϵ environment-mounted

transceivers, and store the measurements in a $\tau\epsilon$ -dimensional measurement vector $Z_k^{U_\tau, E}$. One can also acquire body-relative (intra-user) measurements comprising a single measurement from each of the target user’s τ body-worn transceivers to each of their remaining $\tau - 1$ body-worn transceivers, and form a $\tau(\tau - 1)$ -dimensional measurement vector $Z_k^{U_\tau, U_\tau}$. Finally, one can also acquire cooperative (inter-user) measurements comprising a single measurement from each of the target user’s τ body-worn transceivers to each of the τ body-worn transceivers for each of the $\nu - 1$ reference users and form $\nu - 1$ distinct τ^2 -dimensional measurement vectors $Z_k^{U_\tau, U_{R_i}}$ ($i = 1, 2, \dots, \nu - 1$). The aggregate measurement vector \mathbb{Z}_k includes all environment, body-relative (intra-user), and reference user (inter-user) measurements associated with a particular target user:

$$\mathbb{Z}_k = \left[Z_k^{U_\tau, E}, Z_k^{U_\tau, U_\tau}, Z_k^{U_\tau, U_{R_1}}, Z_k^{U_\tau, U_{R_2}}, \dots, Z_k^{U_\tau, U_{R_{\nu-1}}} \right]^T, \quad (10)$$

and has dimension

$$\begin{aligned} z &= \tau\epsilon + \tau(\tau - 1) + \tau^2(\nu - 1) \\ &= \tau\epsilon + \tau^2\nu - \tau. \end{aligned} \quad (11)$$

The measurement-update of the Kalman filter requires a measurement noise covariance matrix \mathbb{R} , which would have the same dimensionality (rows and columns) as Equation (10). If one assumes there is no correlation between measurements, \mathbb{R} becomes a relatively straightforward diagonal matrix with the diagonal entries set to the expected autocovariance of the range measurements, which could be constant or computed as a function of other conditions.

The theoretical optimality of the Kalman filter assumes the associated random measurement noise is normally-distributed, zero-mean, and spectrally white. Compared to the process noise q magnitude in Equations (7)–(9), the measurement noise magnitude can be estimated with relative ease based on past experience, in a bench-top test setup, or in simulation.

The aggregate Kalman filter measurement update requires both the *actual* measurements \mathbb{Z}_k from Equation (10), and a measurement *prediction* vector $\tilde{\mathbb{Z}}_k$ of the same size. The elements of the measurement prediction vector are computed (not measured) values that indicate what the measurements should be, given the current *a priori* state estimate \mathbb{X}_k^- from Equation (4), and a model for the state-measurement relationships.

In practice the state-measurement model is implemented with a *measurement function* that takes a state vector as input and produces a measurement vector as output. By convention, the function is named h , i.e. a measurement prediction $\tilde{\mathbb{Z}}_k$ would be computed from a state vector X_k via $\tilde{\mathbb{Z}}_k = h(X_k)$. In our case, the measurement function should return the Euclidian distance between *two* transceivers, each represented by its own state vector X_k^1 and X_k^2 . So in our case, the appropriate measurement function would be

$$\tilde{\mathbb{Z}}_k = h(X_k^1, X_k^2), \quad (12)$$

where $\tilde{\mathbb{Z}}_k$ is a scalar distance. Using user-specific state vectors from Equation (2), each which reflects the state of the τ transceivers for that user, we can assemble an aggregate measurement prediction vector to match Equation (10) as

$$\tilde{\mathbb{Z}}_k = \left[h(X_k^{U_T}, X^E)^T, h(X_k^{U_T}, X_k^{U_T})^T, \right. \\ \left. h(X_k^{U_T}, X_k^{U_{R_1}})^T, \dots, h(X_k^{U_T}, X_k^{U_{R_{v-1}}})^T \right]^T, \quad (13)$$

which like \mathbb{Z}_k in Equation (10) has dimension $z = \tau\varepsilon + \tau^2\nu - \tau$ as derived in Equation (11).

The EKF measurement update step also requires the Jacobian representing the partial derivative of the measurement function with respect to each element of estimated state used in the function, i.e. a measurement Jacobian H_k would be computed from the function h and state vector X_k via $H_k = \partial h(X_k) / \partial X_k$. In our case, because the measurement functions compute the Euclidian distance between *two* transceivers as discussed above and shown in Equation (12), the appropriate Jacobian would be

$$H_k = \frac{\partial h(X_k^1, X_k^2)}{\partial X_k^1} + \frac{\partial h(X_k^1, X_k^2)}{\partial X_k^2}, \quad (14)$$

where H_k has the same number of rows as the measurement function h in the numerator and the same number of columns as the state in the denominator. Using user-specific measurement prediction functions from Equation (13) and the two-parameter Jacobian from Equation (14) we can assemble an aggregate Jacobian matrix

$$\mathbb{H}_k = \begin{bmatrix} \left[\frac{\partial h(X_k^{U_T}, X^E)}{\partial X_k^{U_T}} + \frac{\partial h(X_k^{U_T}, X^E)}{\partial X_k^E} \right] \\ \left[\frac{\partial h(X_k^{U_T}, X_k^{U_T})}{\partial X_k^{U_T}} + \frac{\partial h(X_k^{U_T}, X_k^{U_T})}{\partial X_k^{U_T}} \right] \\ \left[\frac{\partial h(X_k^{U_T}, X_k^{U_{R_1}})}{\partial X_k^{U_T}} + \frac{\partial h(X_k^{U_T}, X_k^{U_{R_1}})}{\partial X_k^{U_{R_1}}} \right] \\ \left[\frac{\partial h(X_k^{U_T}, X_k^{U_{R_2}})}{\partial X_k^{U_T}} + \frac{\partial h(X_k^{U_T}, X_k^{U_{R_2}})}{\partial X_k^{U_{R_2}}} \right] \\ \vdots \\ \left[\frac{\partial h(X_k^{U_T}, X_k^{U_{R_{v-1}}})}{\partial X_k^{U_T}} + \frac{\partial h(X_k^{U_T}, X_k^{U_{R_{v-1}}})}{\partial X_k^{U_{R_{v-1}}}} \right] \end{bmatrix}. \quad (15)$$

As most of the measurements do not depend on most elements of the states, the aggregate Jacobian matrix Equation (15) will have a sparse block structure similar to the other aggregate matrices.

Finally, per the normal EKF we compute the aggregate measurement innovation

$$\Delta\mathbb{Z}_k = \mathbb{Z}_k - \tilde{\mathbb{Z}}_k, \quad (16)$$

and the aggregate Kalman gain \mathbb{K}_k as in Equation (1.11) from [WB95], and then compute the a posteriori aggregate state and error covariance estimates corresponding to Equations (1.12) and (1.13) in [WB95]. We then advance the time step and begin the entire predict-correct process over again.

3.3. Measurement Evaluation and Selection

If one was to exhaustively measure all transceiver combinations, the COMOCAP approach could result in a large number of measurements. Here we describe an adaptation of the approach introduced by [HB83] for evaluating and selecting the most valuable measurements at each measurement update step. We refer to this as Measurement-Selection COMOCAP (MS-COMOCAP).

The error covariance matrix \mathbb{P}_k offers an indication of how much confidence the EKF has in its estimates of the state. Because the measurement Jacobian matrix Equation (15) represents the ratio of change in measurement with respect to state, one can use it to project the state error covariance into the measurement space. During the the measurement update step (Section 3.2) we compare this projection to the expected measurement noise to determine which candidate measurements would address the largest uncertainties.

To begin with we note that the real range measurements Z_k in the aggregate measurement vector \mathbb{Z}_k in Equation (10) can be modeled as a *true* (unknown) measurement \hat{Z}_k plus a normally-distributed, zero-mean, and spectrally white random noise signal, i.e.

$$Z_k = \hat{Z}_k + \bar{v}, \quad (17)$$

where $\bar{v} \sim \mathcal{N}(0, R)$ is an appropriately-sized vector of random variables representing the real measurement noise, and R is the same covariance discussed after Equation (10) above. We zero out any elements of $\Delta\mathbb{Z}_k$ corresponding to *unchosen* measurements, to eliminate any effects on the corresponding state elements.

To simplify the remaining explanation we eliminate the time step k notation. Considering the aggregate versions of the vectors and matrices, including an appropriately sized aggregate measurement noise vector \bar{V} corresponding to \bar{v} above, and noting from Equation (17) that $\bar{v} = Z - \hat{Z}$, we formulate $\tilde{\mathbb{Z}}$ as the ratio of measurement prediction error vector to measurement noise:

$$\tilde{\mathbb{Z}} = (\tilde{\mathbb{Z}} - \hat{\mathbb{Z}}) / (Z - \hat{Z}) \\ = (\tilde{\mathbb{Z}} - \hat{\mathbb{Z}}) / \bar{V}. \quad (18)$$

Note that like \mathbb{Z}_k in Equation (10) and $\tilde{\mathbb{Z}}_k$ in Equation (13), $\tilde{\mathbb{Z}}$ in Equation (18) has z elements. As such we can define the elements of $\tilde{\mathbb{Z}}$ as $[\tilde{Z}_1, \tilde{Z}_2, \dots, \tilde{Z}_z]$, and consider a weighted combination of the measurement prediction error elements:

$$\mathbb{W} = a_1\tilde{Z}_1 + a_2\tilde{Z}_2 + \dots + a_z\tilde{Z}_z \\ = a^T\tilde{\mathbb{Z}} \quad (19)$$

we can then reformulate \mathbb{W} as an error covariance

$$\sigma_W^2 = \sum_{i=1}^z \sum_{j=1}^z a_i a_j (\mathbb{S}_{i,j} / \mathbb{R}_{i,j}) = a^T (\mathbb{S} / \mathbb{R}) a, \quad (20)$$

where $\mathbb{S}_{i,j}$ are the individual elements of

$$\mathbb{S} = \mathbb{H}\mathbb{P}^{-1}\mathbb{H}^T, \quad (21)$$

which is the state error covariance projected into the measurement space, and $\mathbb{R}_{i,j}$ are the individual elements of the measurement noise covariance \mathbb{R} described after Equation (10).

A relatively large σ_W^2 in Equation (20) would indicate a relatively large reduction of state error covariance (state estimation uncertainty) for the corresponding measurement. As such we seek a linear combination of measurement choices, based on the current statistics and models, that maximizes σ_W^2 under the constraint that $a^T a = 1$. To find the optimal weightings we use the Lagrangian multiplier method:

$$\frac{\partial[\sigma_W^2 - \lambda(a^T a - 1)]}{\partial a} = 0. \quad (22)$$

By substituting Equation (20) into Equation (22), we get

$$\frac{\partial [a^T (\mathbb{S}./\mathbb{R})a - \lambda(a^T a - 1)]}{\partial a} = 0, \quad (23)$$

which simplifies to

$$[(\mathbb{S}./\mathbb{R}) - \lambda]a = 0. \quad (24)$$

In general the solution to Equation (24) can be found by determining the eigenvectors and eigenvalues of $\mathbb{S}./\mathbb{R}$. Recall that we assume every measurement is independent, therefore \mathbb{R} is a diagonal matrix. As such one can choose measurements by simply dividing the eigenvalues of \mathbb{S} by the diagonal elements of \mathbb{R} , element by element, to obtain a series of ratios that indicate the impact of the corresponding measurement on the state error covariance. The larger the ratio, the greater that measurement will impact on the state estimation. As such we sort all of the ratios in descending order, and obtain and use the top measurements for the Kalman filter measurement update. For our simulation setup (described below) we used the top 2/3 measurement contenders.

4. Pilot Experiments

Here we describe the pilot simulation experiments we performed to evaluate and compare the three methods: COMOCAP EKF, MS-COMOCAP EKF, and a standard EKF implementation.

4.1. Materials

We based our simulation on a real optical MOCAP setup in our lab, with six OptiTrack cameras mounted on three tripods and arranged around a small room-sized real walking area. As shown in Figure 2 we used the OptiTrack system (left) to track three points on the user’s body: head, left wrist, and right wrist; and we simulated 13 body-mounted transceivers (right) on each of two humans, plus two environment-mounted transceivers on two of the three tripods.

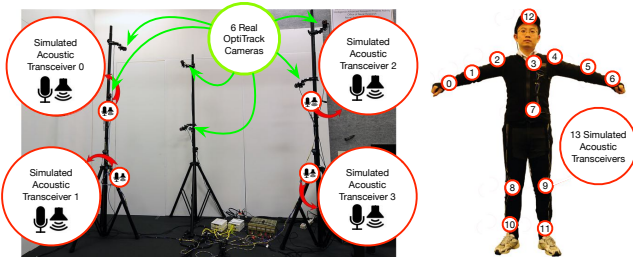


Figure 2: Illustration of our simulation setup: Left: three tripods with each two OptiTrack cameras and two environment-mounted transceivers on two tripods. Right: three tracked OptiTrack markers (head, left wrist, right wrist) and 13 body-mounted transceivers.

4.2. Method

Movement Scenario. We based our simulation on real MOCAP data collected from two people who walked towards each other over a distance of approximately two meters, and then shook their hands. This scenario allowed us to compare inter-individual distances between two people, leading up to the worst-case situation of existing MOCAP systems that occurs when two people touch each other.

Ground Truth Data. We captured the three body points over the scenario, filtered the data, and treated the smooth tracks as the “ground truth” for our simulations. Capturing data simultaneously from two actors performing turned out to be problematic due to missing (occluded) or inaccurate tracking data at the moment when they shook their hands. As such we decided to capture the movements of the two actors separately, and then combine them.

Simulation. Our simulation was performed in *Matlab*. We calculated the distances between every two transceivers, including *inter-individual* distances between transceivers on each participant, *intra-individual* distances between each two transceivers on one participant, and distances between the body-mounted and environment-mounted transceivers. We simulated measurement noise by adding a normally distributed zero mean signal to the distances. Based on published noise magnitudes from [FHP98, Val02] we used $\sigma = 2$ mm. We used this same magnitude for the measurement covariance matrix R of the Kalman filter. For our COMOCAP EKF method and MS-COMOCAP EKF method, we acquired both the cooperative measurements and environment-reference measurements. For the original EKF method, we only simulated the environment-reference measurements. We simulated occlusions by excluding three of the four *environment*-reference measurements for the entire simulation.

4.3. Results and Discussion

Wrist Motion. Figure 3(left) shows the x position of one actor’s right wrist along the main movement direction towards the other actor and its estimation by the three considered methods (COMOCAP EKF, MS-COMOCAP EKF, and original EKF). Figure 3(right) shows a zoomed-in view of the dashed-line window—the short window from $frame = 100$ to $frame = 200$. Overall our COMOCAP EKF estimation (in red) is the closest to the ground truth (black) while the original EKF estimation (blue) is the furthest from the ground truth. The MS-COMOCAP EKF estimation (green) is in between. These overall results may be explained by the information used by each method: the COMOCAP EKF used the cooperative and environment-reference measurements, while the original EKF used only the environment-reference measurements. For MS-COMOCAP EKF, the top 2/3 of the measurements were used, including cooperative and environment-reference measurements. As can be seen in Figure 3(right), the period associated with frames 110–130 exhibits relatively large error for all of the methods. This is a result of EKF prediction overshoot given our filter tuning (EKF model parameter settings) and the prolonged period of relatively constant velocity during the period associated with frames 90–110 shown in Figure 3. The use of a multi-modal (multiple model) filter approach [ABW01] would likely improve this situation.

To help illustrate the performance we plotted each method’s estimation errors, i.e., the differences between the estimates and the ground truth. As shown in Figure 4(left), the COMOCAP EKF estimates are closest to the ground truth. Although MS-COMOCAP EKF took only 2/3 the measurements of COMOCAP EKF, it performed only a little worse. The original EKF performed worst as its estimation is the furthest from the ground truth.

Shaking Hands. The closer the two participants get, the more accurate their *relative* positions need to be. With the COMOCAP

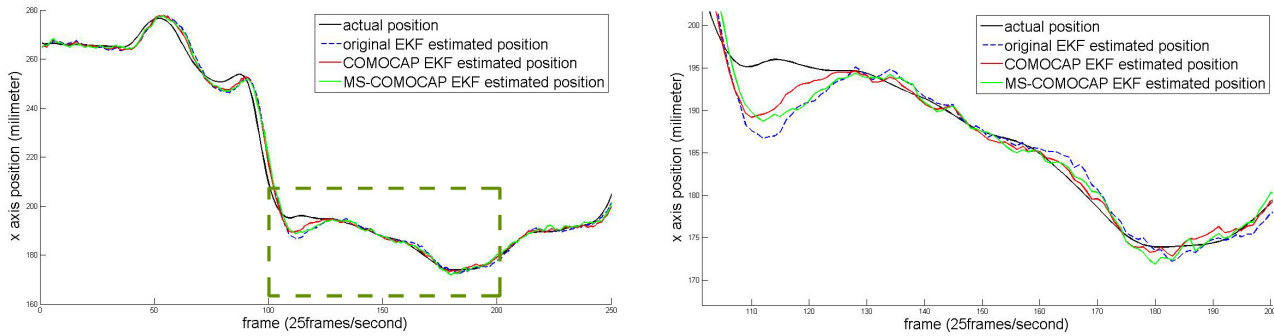


Figure 3: (Left) The ground truth data of one participant's right wrist along the x-axis position and its estimations by COMOCAP EKF, MS-COMOCAP EKF, and the original EKF. (Right) Zoom-in view of the dashed-line window.

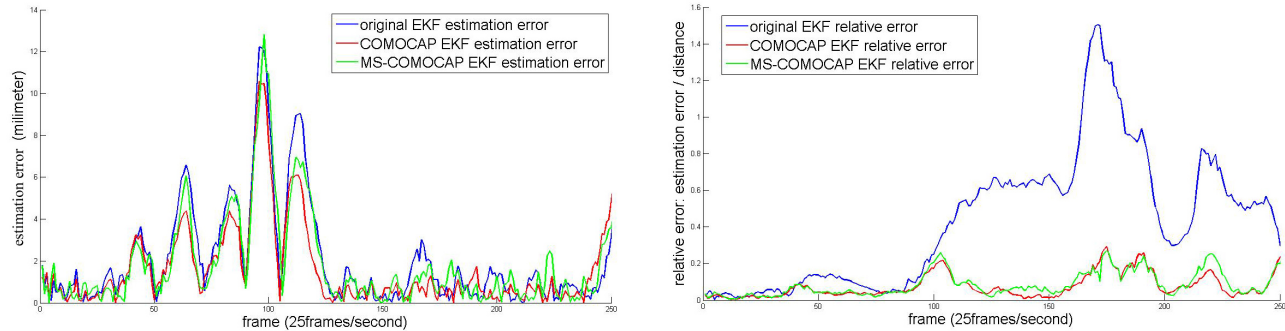


Figure 4: (Left) Errors in COMOCAP EKF, MS-COMOCAP EKF and the original EKF estimation compared to the ground truth. (Right) The relative errors of COMOCAP EKF, MS-COMOCAP EKF and the original EKF estimation when the distance between the two wrist points decreased (proximity scenario).

methods, the cooperative measurements between the two participants provided a direct observation of their relative position. Since the two participants got close and shook their hands with their right hands in the simulation experiment, we took both participants' right wrist points and plotted the relative error of the two points' 3D estimation error covariances summation over their distance. As visualized in Figure 4(right), when the two wrist points get closer to each other in order to shake their hands, the relative errors of both COMOCAP EKF and MS-COMOCAP EKF successfully remained at similar levels. However, for the original EKF estimation, the relative error increased greatly due to the lack of direct observation data of their relative positions.

Occlusion Case. We evaluated situations when three of the four environment-fixed reference points were occluded (see Figure 5). The figure shows that even though there was only one environment-reference measurement, both COMOCAP EKF and MS-COMOCAP EKF were still able to estimate motions because the cooperative measurements between participants provided enough observations. However, the original EKF, which only considered environment-reference measurements, failed to estimate motions because there were not enough observations acquired.

Overall, our pilot simulation results suggest that both COMOCAP EKF and MS-COMOCAP EKF can provide improvements over the original EKF method in three ways:

1. the accuracy of absolute position estimates can be increased due to the additional cooperative measurements;
2. the relative errors between two users can be reduced due to the direct measurements between sensors on their bodies, which are more pronounced when the sensors are getting very close; and

3. the robustness to occlusion can be greatly improved when two users are close to each other compared to classical environment-reference MOCAP tracking.

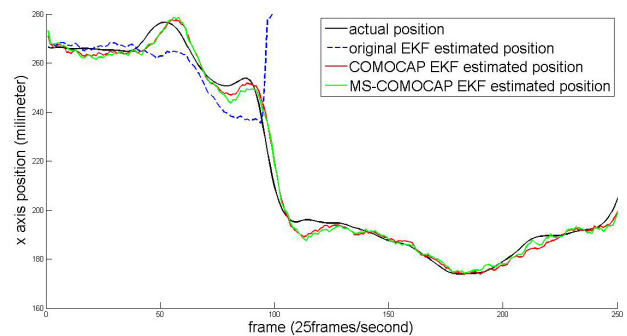


Figure 5: Estimations of COMOCAP EKF, MS-COMOCAP EKF and the original EKF when three of the four environment reference points were occluded (occlusion scenario).

5. Conclusion and Future Work

We presented a novel approach to MOCAP that combines and extends conventional global and body-relative approaches by distributing both sensing and active signaling over each person's body to facilitate body-relative (intra-user) and *body-body* (*inter-user*) measurements for multiple people, in an approach we call *cooperative motion capture* (COMOCAP). Simulation results from a COMOCAP system comprised of acoustic *transceivers* suggest advantages in terms of improving accuracy and robustness to occlusions in situations of close proximity between multiple persons. COMOCAP could improve existing and novel tracking systems.

Related to our work, Vallidis and Bishop presented an acoustic ranging approach that, unlike narrow band (e.g., ultrasonic) acoustic systems, is relatively robust to occlusions as the spread spectrum signals can diffract around objects and still estimate distance [Val02]. However acoustic sensors are not omnidirectional—signal strength is dependent on angle in comparison to retroreflective optical markers. We believe that the most robust approach would be to combine multiple modalities, e.g., optical tracking for the environment references and acoustic ranging for body-relative and cooperative measurements. The EKF-based approach is general enough to support any hybrid combination of modalities.

In the future we plan to extend it to a hybrid optical-acoustic system, e.g., by combining it with HTC's Lighthouse 2 in confined physical spaces (*Room-Scale VR*), where we expect inter- and intra-user occlusions as discussed in this paper to be prevalent.

6. Acknowledgements

The work presented in this publication was supported in part by the Office of Naval Research (ONR) Code 30 under Program Officers Dr. Peter Squire (ONR awards N00014-14-1-0248, N00014-12-1-1003, and N00014-12-1-0052) and Dr. Roy Stripling (N00014-08-C-0349). We also acknowledge Florida Hospital for their support of Prof. Welch via their Endowed Chair in Healthcare Simulation.

References

- [ABW01] ALLEN B. D., BISHOP G., WELCH G.: Tracking: Beyond 15 minutes of thought. *ACM SIGGRAPH Course*. 2001. 3, 6
- [Ani17] ANIMAZOO: Gypsy7, <http://metamotion.com/gypsy>, 2017. accessed 30 June 2017. 2
- [ART17] Advanced realtime tracking artrack, <http://www.ar-tracking.com>, 2017. accessed 30 June 2017. 1, 2
- [Atr17] ATRACSYS: accuTrack500, <https://atracsys.com/web/eng/measurement>, 2017. accessed 30 June 2017. 2
- [BH96] BROWN R. G., HWANG P. Y. C.: *Introduction to Random Signals and Applied Kalman Filtering: with MATLAB Exercises and Solutions*, third ed. Wiley & Sons, Inc., 1996. 3, 4
- [DB03] DOBRIAN C., BEVILACQUA F.: Gestural control of music: using the vicon 8 motion capture system. In *New Interfaces for Musical Expression* (2003), pp. 161–163. 2
- [DOKA13] DAMIAN I., OBAID M., KISTLER F., ANDRÉ E.: Augmented reality using a 3D motion capturing suit. In *ACM Augmented Human* (2013), pp. 233–234. 1, 2
- [DU02] DORFMÜLLER-ULHAAS K.: *Optical Tracking From User Motion To 3D Interaction*. PhD thesis, Vienna University of Technology, 2002. 2
- [FCOL99] FLEISHMAN S., COHEN-OR D., LISCHINSKI D.: Automatic camera placement for image-based modeling. In *Pacific Conference on Computer Graphics and Applications* (1999). 2
- [FHP98] FOXLIN E., HARRINGTON M., PFEIFER G.: Constellation: A wide-range wireless motion-tracking system for augmented reality and virtual set applications. In *Computer Graphics*, Cohen M. F., (Ed.), SIGGRAPH 98 conference proceedings ed., Annual Conference on Computer Graphics & Interactive Techniques. ACM Press, Addison-Wesley, Orlando, FL USA, 1998, pp. 371–378. 6
- [HB83] HAM F. M., BROWN R. G.: Observability, eigenvalues, and kalman filtering. *IEEE Trans. Aerosp. Electron. Syst* (1983), 269–273. 5
- [HKS*15] HE C., KAZANZIDES P., SEN H. T., KIM S., LIU Y.: An inertial and optical sensor fusion approach for six degree-of-freedom pose estimation. *Sensors* 15, 7 (2015), 16448–16465. 2
- [HSGS06] HOL J., SCHÖN T., GUSTAFSSON F., SLYCKE P.: Sensor fusion for augmented reality. In *Proceedings of the International Conference on Information Fusion* (2006). 2
- [JSG15] JOHANNSEN O., SULC A., GOLDBLUECKE B.: On linear structure from motion for light field cameras. In *IEEE International Conference on Computer Vision* (2015), pp. 720–728. 2
- [JWF*09] JOHNSON T., WELCH G., FUCHS H., LA FORCE E., TOWLES H.: A distributed cooperative framework for continuous multi-projector pose estimation. In *IEEE VR* (2009), pp. 35–42. 3
- [Kal60] KALMAN R. E.: A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering* 82, Series D (1960), 35–45. 3
- [KPP*13] KRIGSLUND R., POPOVSKI P., PEDERSEN G. F., DIDERIKSEN J. L., FARINA D., DOSEN S.: A novel technology for motion capture using passive UHF RFID tags. *IEEE Transactions on Biomedical Engineering* 60, 5 (2013), 1453–1457. 2
- [Lok02] LOK B.: *Interacting with Dynamic Real Objects in Virtual Environments*. PhD thesis, Department of Computer Science, University of North Carolina at Chapel Hill, 2002. 2
- [MHK06] MOESLUND T. B., HILTON A., KRÜGER V.: A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104 (2006), 90–126. 1, 2
- [Neu17] Perception neuron, <https://neuronmocap.com>, 2017. accessed 30 June 2017. 2
- [Not] Notch, <https://wearnotch.com>. accessed 30 June 2017. 1
- [Opt17a] OptiTrack Prime41, <http://optitrack.com/products/prime-41/>, 2017. accessed 30 June 2017. 1, 2
- [Opt17b] Optotrak certus, <https://www.ndigital.com/msci/>, 2017. accessed 30 June 2017. 1, 2
- [Org17] Organic motion markerless mocap, <http://www.organicmotion.com>, 2017. accessed 30 June 2017. 2
- [Osh06] OSHITA M.: Motion-capture-based avatar control framework in third-person view virtual environments. In *ACM SIGCHI Int. Conf. on Advances in Computer Entertainment Technology* (2006), p. 2. 2
- [Pha17] PHASESPACE: Impulse x2e, <http://phasespace.com/x2e-motion-capture/>, 2017. accessed 30 June 2017. 1, 2
- [Pol17] Polhemus liberty latus, <http://polhemus.com>, 2017. accessed 30 June 2017. 1, 2
- [RDW91] RAO B. S., DURRANT-WHYTE H. F.: Fully decentralised algorithm for multisensor kalman filtering. *IEE Proceedings D - Control Theory and Applications* 138, 5 (1991), 413–420. 3
- [RK17] RAHIMIAN P., KEARNEY J. K.: Optimal camera placement for motion capture systems. *IEEE Trans. Vis. Comput. Graph* 23, 3 (2017), 1209–1221. 2
- [RKS*05] ROSENHAHN B., KERSTING U. G., SMITH A. W., GURNEY J. K., BROX T., KLETTE R.: A system for marker-less human motion estimation. In *Joint Pattern Recognition* (2005), pp. 230–237. 2
- [Val02] VALLIDIS N. M.: Whisper: A spread spectrum approach to occlusion in acoustic tracking, 2002. 2, 6, 8
- [VAV*07] VLASIC D., ADELSBERGER R., VANNUCCI G., BARNWELL J., GROSS M., MATUSIK W., POPOVIĆ J.: Practical motion capture in everyday surroundings. In *ACM SIGGRAPH Papers* (2007). 2
- [VIC] Vicon, <https://www.vicon.com>. acc. 30 June 2017. 1, 2
- [WB95] WELCH G., BISHOP G.: *An Introduction to the Kalman Filter*. Tech. Rep. TR95-041, University of North Carolina at Chapel Hill, Department of Computer Science, 1995. 3, 5
- [WB01] WELCH G., BISHOP G.: An introduction to the kalman filter. *ACM SIGGRAPH Course*. 2001. 4
- [WF02] WELCH G., FOXLIN E.: Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Comput. Graph. Appl.* 22, 6 (2002), 24–38. 1, 2
- [ZP17] ZIHAJEZHDADEH S., PARK E. J.: A novel biomechanical model-aided imu/uwb fusion for magnetometer-free lower body motion capture. *IEEE Trans Syst Man Cybern Syst* 47, 6 (2017), 927–938. 2