Estimating Human Dynamics On-the-fly Using Monocular Video For Pose Estimation

Priyanshu Agarwal[†], Suren Kumar[†], Julian Ryde[‡], Jason J. Corso[‡] and Venkat N. Krovi[†]
Mechanical and Aerospace Engineering Department[†]
Department of Computer Science and Engineering[‡]
State University of New York at Buffalo
Buffalo, New York 14260-1660
Email: {priyansh, surenkum, jryde, jcorso, vkrovi}@buffalo.edu

***Abstract*—Human pose estimation using uncalibrated monocular visual inputs alone is a challenging problem for both the computer vision and robotics communities. From the robotics perspective, the challenge here is one of pose estimation of a multiply-articulated system of bodies using a single non-specialized environmental sensor (the camera) and thereby, creating low-order surrogate computational models for analysis and control.**

In this work, we propose a technique for estimating the lower-limb dynamics of a human solely based on captured behavior using an uncalibrated monocular video camera. We leverage our previously developed framework for human pose estimation to (i) deduce the correct sequence of temporally coherent gap-filled pose estimates, (ii) estimate physical parameters, employing a dynamics model incorporating the anthropometric constraints, and (iii) filter out the optimized gap-filled pose estimates, using an Unscented Kalman Filter (UKF) with the estimated dynamically-equivalent human dynamics model. We test the framework on videos from the publicly available DARPA Mind's Eye Year 1 corpus [8]. The combined estimation and filtering framework not only results in more accurate physically plausible pose estimates, but also provides pose estimates for frames, where the original human pose estimation framework failed to provide one.

I. Introduction

Estimating and tracking 3D pose of humans in unconstrained environments using monocular vision poses several technical challenges due to high-dimensionality of human pose, self-occlusion, unconstrained motions, variability in human motion and appearance, observation ambiguities (left/right limb ambiguity), ambiguities due to camera viewpoint, motion blur and unconstrained lighting [13]. In the past, visual tracking algorithms have relied on kinematic prior models. While these exploit the motion constraints of articulated models, joint limits, and temporal smoothness for improved performance, they have proven insufficient to capture the nonlinearities of human motion [28, 27, 14]. More advanced activity-specific models learned from motion capture data have also been employed that yield approximately correct body configurations on datasets containing similar motions [9, 29, 30]. However, such models often lead to poses that are physically implausible (e.g. foot-skates, out-of-plane rotations) and face difficulty in generalization beyond the learned dataset.

Brubaker et al. [6] introduced the notion of a physics-based biomechanical prior-model characterizing lower-body dynamics. A key benefit is the natural accommodation of variations in style due to changes in speed, step-length, and mass leading to 3D person tracking with physically plausible poses. The Kneed Walker extension, allowed for capture of subtle aspects of motion such as knee bend and torso lean, even when these are not strongly evident from the images [5]. Vondrak et al. [31] employed a full body 3D simulation-based prior that explicitly incorporated motion control and dynamics into a Bayesian filtering framework for human motion tracking. However, there is always one or more fundamental assumptions involved that there is a priori knowledge about the physical properties (e.g. mass, inertia, limb lengths, ground plane and/or collision geometries), the activity in the scene, calibrated camera, imagery from multiple cameras, availability of similar motion dataset [24, 2, 3, 25, 16, 26, 32]. Furthermore, the obtained pose estimates are susceptible to the error in the physical parameters estimates of the system due to propagation of pose states via a nonlinear dynamics model.

Brubaker et al. [4] proposed a framework for estimating contact dynamics and internal torques using 3D pose estimates obtained from two views (roughly sagittal and roughly frontal) of a subject using calibrated cameras, mass and inertial properties determined by combining the body segment length estimated from the motion capture data with standard biomechanical data. However, well established system identification techniques fail to solve the problem of estimating physical human parameters solely using uncalibrated monocular imagery when only partial noisy pose estimates are available due to: (i) partial state knowledge (no pose velocity information); (ii) noise due to imaging inaccuracies; (iii) unknown and insufficient input excitation (both structure and frequency); (iv) low sampling frequency; and (v) inherent dynamics nonlinearity. Nevertheless, accurate and efficient kinematic and inertial parameter estimation of articulated multibody systems using noisy partial sensing has broader applications for enhanced telepresence [22], robot navigation [12], vision-based humanoid control, multi-robot cooperation [20] and imitation based robot control [21].

In this work, we build on our previous work [1] wherein a framework for human pose estimation in uncalibrated monocular videos is proposed to obtain the raw pose estimates of lower limbs. A Fast Fourier Transform (FFT) of the raw pose estimates provides information regarding the fundamental fre-

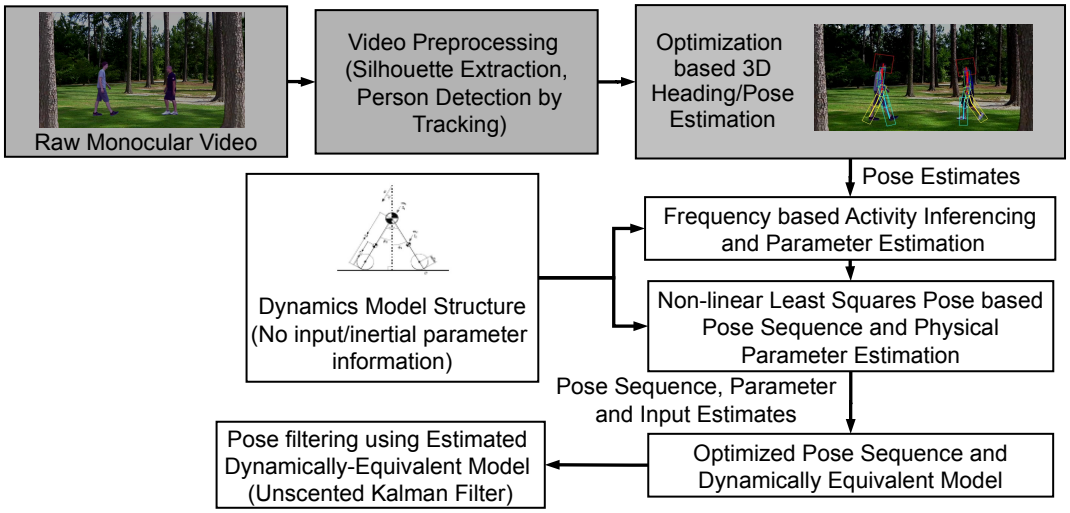


Fig. 1: Overview of the overall system. The grayed out boxes represent previous work of the authors that has been used in this framework.

quency in the estimates leading to inference about the current activity in the scene. We focus on inertial and kinematic parameter estimation for human lower-limbs from monocular video by setting up optimization subproblems employing the structure of the dynamics model. A simplified dynamics model–the Anthropomorphic Walker is used along with the optimized anthropometric constraints tuned to generate stable gait, to identify the human physical parameters along with an optimized sequence of gap-filled pose estimates. Finally, an UKF is used to filter out the optimized gap-filled pose estimates using the continuous-discrete state-space model of the estimated dynamically-equivalent human dynamics model for walking.

II. System Overview

Problem Statement: Given the noisy partial pose estimates (joint angles of hips) from an uncalibrated monocular video of a human, develop a dynamic (equivalent) model for human activity suitable for subsequent use to filter the pose estimates.

Fig. 1 provides an overview of the proposed system. A raw monocular video is processed to extract human silhouette and track humans in the scene. The human pose estimation framework is then used to obtain the pose estimates. The lower limb pose information is used to extract the primary frequency component and inference about the activity in the scene. An anthropometrically constrained dynamic model with uncertain physical parameters is simulated to obtain a model-based frequency estimate. In Phase I, a non-linear least squares (NLS) estimator is used to estimate a subset of the parameters that minimizes the frequency disparity (between the measured and simulated frequencies).This is used to bootstrap a Phase II NLS estimator (with 6 uncertain parameters) to minimize the disparity between measured and simulated poses. These estimates are then used to obtain a dynamically equivalent system which is used as the process model for an UKF to obtain the filtered pose estimates.

III. Human Pose Estimation

The work on human pose estimation by the authors employs a model-based generative approach for the task of human pose estimation in unrestricted environments. Unlike many previous approaches, the framework is fully automatic, without any manual initialization for monocular video-based pose estimation. We do not use any camera calibration, prior motion (motion capture database), prior activity, appearance, body shape or size information about the scene. The generative model for estimating heading direction of the subject in the video uses motion-based cues to constrain the pose search space. Thereafter, a sequential optimization framework estimates the remaining uncoupled pose states (camera parameters, body location, body joint angles) leveraging a combination of deterministic and probabilistic optimization approaches. The framework outputs two probable body pose estimates (human torso location (x, y, scaled z), right/left hip flexion angle, right/left knee flexion angle, human heading direction) for each frame in the video. Fig.s. 7(a), 7(b) illustrates the two obtained raw pose estimates (right leg being forward and the left leg being forward, respectively) obtained for a few videos in the DARPA corpus [8]. However, determining the correct temporally coherent sequence of physically plausible pose estimates remained a challenge which is addressed in this work for nearly periodic activities.

IV. Human Dynamics Model

A. The Anthropomorphic Walker

The Anthropomorphic Walker proposed and pioneered by McGeer [19], and later modified by Kuo [17] by introducing impulsive push-off combined with torsional spring between the legs permitted the generation of a range of stable walking

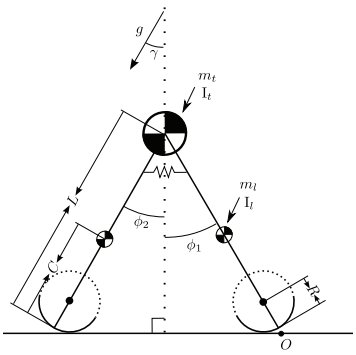


Fig. 2: Parameters of the Anthropomorphic Walker. Please refer to [6] for details regarding the dynamics model.

motion on level and slanted ground. We use this simplified dynamics model (Eqn.s 1,2) for modeling human walking.

$$\mathbf{T}^T M \mathbf{T} \ddot{q} = f(t) + \mathbf{T}^T M (G - \mathbf{g}) \tag{1}$$

where $q = [\phi_1, \phi_2]^T$, $T(q)$ is the kinematic transfer function, M is the mass matrix, $\mathbf{T}$ is the Jacobian of $T(q)$, $f(t)$ is the generalized force vector, G is the gravitational acceleration vector, $\mathbf{g}$ is the convective acceleration, and γ is the ground slope angle which is assumed to be zero for this work. The post collision velocities are solved using

$$\mathbf{T}^{+T} M \mathbf{T}^{+} \dot{q}^{+} = \mathbf{T}^{+T} \left(S + M \mathbf{T} \dot{q}^{-} \right) \tag{2}$$

where $T^{+}(q)$ is the kinematic transfer matrix immediately after the impact, $\mathbf{T}^{+}(q)$ is the Jacobian of $T^{+}(q)$, $\dot{q}^{-}$, $\dot{q}^{+}$ are the pre- and post-collision velocities, respectively, and S is the impulse vector.

B. Incorporation of Empirical Anthropomorphic Constraints

In order to constrain the various physical parameters (lengths, mass, inertia) in the original model, we exploit the empirical anthropometric values provided in [7]. However, since the dynamics model is a simplified version of the actual human anthropometry, the values are not directly applicable. We set up an optimization problem (Eqn. 3) to estimate the optimal length relations such that minimum variation is obtained in step length across multiple steps (i.e. gait is stable) by simulating the dynamics model for the current set of anthropometric relations. We initialize these relations with the corresponding relations in standard human anthropometry.

$$\begin{aligned} \min_{\Theta_a} \quad & f(\Theta_a) = Var(l_{si}) \\ \text{subject to: } & \Theta_{al}^{(j)} \leq \Theta_a^{(j)} \leq \Theta_{au}^{(j)}, l_{si} > 0, n_s > n_{min} \\ & \Theta_a = [r_R, r_C, r_s, r_{th}, r_{\gamma_l}, r_{\gamma_t}] \end{aligned} \tag{3}$$

where Var(.) stands for variance, $l_{si}, r_x, r_{\gamma_x}, n_s, n_{min}$ represent the step length of i^{th} step, relative fraction of the parameter with respect to the total height of the human, relative fraction of radius of gyration with respect to the length of the body part, the number of steps, and the minimum number of foot steps required in simulation. We use $\Theta_{al} = [0.1, 0.6, 0.2, 0.2, 0.3, 0.4]$ and $\Theta_{au} = [0.2, 0.7, 0.3, 0.3, 0.4, 0.5]$ such that the standard relations are well within the bounds. In order to obtain global solutions in the multi-modal space, we use the standard genetic algorithm [11] to solve the optimization problem with termination criteria as tolerance on parameter values (0.001) and function value (0.01). The optimized length relationships along with other mass relationships are provided in Table I. The only unknowns for the model after incorporating these constraints are the total mass, total height, spring stiffness for the spring connecting the two legs, and the magnitude of foot-ground collision impulse of the human. However, since the spring stiffness and the magnitude of foot-ground collision impulse leads to the same effect in the dynamics, we treat the impulsive reaction force to be a constant. Please note that till this point no pose information regarding the specific person in the scene has been exploited.

Mass Relations		
Body Part	**Notation**	**Relative fraction/ Expression**
Torso/Upper body	m_t	0.678
Thigh	m_{th}	0.1
Shank	m_s	0.061
Leg	m_l	$m_{th} + m_s$
Length Relations		
Upper body	l_t	0.47
Thigh	l_{th}	0.23669
Shank	l_s	0.24556
Foot Radius	R_f	0.14466
Center of Mass Relations		
Leg from foot end	C_l	0.553
Radius of Gyration Relations		
Leg	γ_l	0.326
Torso	γ_t	0.496
Inertia Relations		
Leg	I_l	$m_l \times \gamma_{leg}^2$
Torso	I_t	$m_t \times \gamma_t^2$

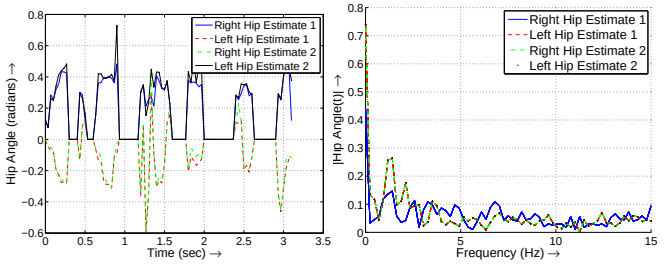
TABLE I: Relative fractions and expressions obtained after optimizing the anthropometric relationships for various physical parameters. Relative fractions are expressed in terms of full body mass and length.

V. Pose and Parameter Estimation

Once a stable walking model is obtained, the parameters can now be optimized to generate the characteristic nearly periodic gait of the human in the scene using the raw pose estimates in a two stage approach.

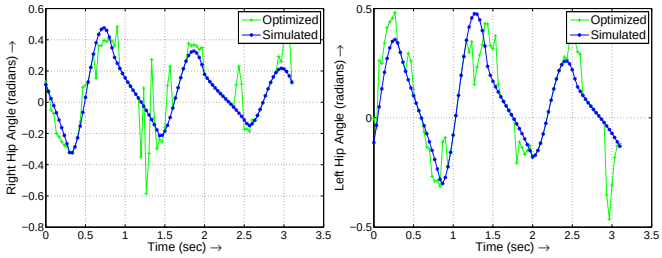
A. Gait Frequency Based Estimation

Estimation of all parameters simultaneously based on pose alone leads to incorrect estimates at times due to aliasing. Physical parameter estimation to first minimize the difference



(a) **(b)**

Fig. 3: Frequency estimation from raw pose estimates for a video in the DARPA corpus. (a) Time evolution of raw pose estimates, and (b) Single-sided amplitude spectrum of raw pose estimates.



(a) **(b)**

Fig. 4: Optimized gap-filled and dynamics-simulation-based pose estimates for a video in the DARPA corpus. (a) Right hip angle, and (b) Left hip angle.

between the simulated and estimated gait frequency reduces this error. Fig. 3 represents the raw pose estimates and the single-sided amplitude spectrum of the raw pose estimates obtained using FFT. We take the average of the frequency estimates obtained using the right (f_{rh1}, f_{rh2}) and the left (f_{lh1}, f_{lh2}) hip raw pose estimates for both the raw estimates (Eqn. 4). The range in which the frequency of the raw pose estimates lies determines the periodic activity of the human in the scene. Fujiyoshi and Lipton [10] have shown that it is possible to classify motions into walking and running based on the frequency of the cyclic motion and that a threshold of 2.0 Hz correctly classifies 97.5% of the target motions. In this work, we consider the case where the human activity is walking ($f_e \approx 1.17Hz$) to illustrate the concept. Generalized more complex dynamics model can be built that can cater multiple nearly periodic activities. In this Phase I NLS estimation problem (Eqn. 5), we seek to minimize the difference between estimated frequency (f_e) and the frequency obtained using the dynamics simulation (f_s).

$$f_e = \frac{f_{rh1} + f_{lh1} + f_{rh2} + f_{lh2}}{4} \tag{4}$$

$$\begin{aligned} \min_{\Theta} \quad & f(\Theta) = \|f_s - f_e\| \\ \text{subject to: } & \Theta_l \leq \Theta \leq \Theta_u, \Theta = [m, h, \kappa] \end{aligned} \tag{5}$$

We use $\Theta_l = [1, 0.1, 0.1]$ and $\Theta_u = [200, 2, 100]$ for the optimization problem to be generalizable.

B. Pose Based Estimation

In Phase II, we tune the full parameter set of the model to minimize the difference between the dynamics-simulation-based states and the estimated states (chosen based on its proximity to the simulated state for the current frame). Initializing the dynamics model with the noisy raw pose states also results in unstable walking (due to noise in the estimation). So, the initial states are also estimated during the parameter estimation. The objective is to minimize the difference between the dynamics-simulation based states and the estimated raw states (Eqn. 6).

$$\begin{aligned} \min_{\Theta} \quad & f(\Theta) = \sum_{j=1}^{N} \min_{i} \|X_s - X_{ei}\|_j \\ \text{subject to: } & \Theta_l \leq \Theta \leq \Theta_u, \\ & \Theta = [m, h, \kappa, \phi_{10}, \phi_{20}, \dot{\phi}_{10}, \dot{\phi}_{20}], \end{aligned} \tag{6}$$

where $X = [\phi_1, \phi_2]$, X_s and X_{ei}, $i = 1, 2$ represent the simulated and the two estimated poses, respectively, and Θ represents the unknown parameters to be estimated. Furthermore, experiments showed that assuming $\phi_{20} = -\phi_{10}$ resulted in stable walking gait which otherwise was difficult. We use $\Theta_l = [1, 0.1, 0.1, -\pi/2, -\pi/2, -3\pi/2, -3\pi/2]$ and $\Theta_u = [200, 2, 100, \pi/2, \pi/2, 3\pi/2, 3\pi/2]$. We use the function evaluation based Nelder-Mead Simplex Method (due to absence of closed form solution of the states making derivative-based methods inapplicable) [18] to solve both the optimization problems with termination criteria as tolerance on parameter values (0.01), function value (0.01) and maximum number of function evaluations (100000).

During the optimization for pose and physical parameters we obtain (i) the correct sequence of the pose estimates as well as (ii) the estimates for the physical parameters of the system resulting in a dynamically-equivalent model for the human in the scene. Raw pose estimates for many frames are missing due to inability of the pose estimation algorithm to provide reliable estimates for these frames because of: (i) improper silhouette segmentation due to insignificant lower limb movement, large shadow, high similarity in appearance of foreground and background, merging of human silhouette with other entities; (ii) partial occlusion of the human in the scene. We substitute the dynamic simulation based states for the states available from simulation (hip angles) and use linear interpolation for the remaining states. Fig. 4 shows the optimized gap-filled and dynamics simulation-based pose estimates for the hip angles in a video.

In order to filter the optimized pose estimates using the estimated dynamically-equivalent model we linearized the dynamics model. However, the linearization of the model does not truly capture the non-linearity in the model especially because a strong non-linearity is present in the system (due

to impulsive collision of the foot with the ground). This leads to unstable walking, rendering the use of Extended Kalman Filter infeasible. Hence, in lieu of computationally expensive particle filters, we explore the use of UKF for filtering out the optimized states, using the dynamically equivalent model for the human in the scene.

VI. Unscented Kalman Filtering

The original discrete-time Unscented Kalman filter works [15] satisfactorily when small time steps are considered. However, for the problem of human pose filtering the measurement update rate (Δt) is governed by the frame rate of the video being processed. This is typically poor even for high frame-rate cameras (~ 100 Hz), especially given significant non-linearity in the dynamics. Furthermore, the presence of discontinuity in the system dynamics requires a continuous-discrete Kalman filter to be employed, especially because the occurrence of the discontinuity significantly affects the stability of the system. The use of continuous time non-linear state-space model for the system ensures that the occurrence of the discontinuity is accurately captured in time, which is imperative for the stability of the system. In addition, the resulting change in the system model requires careful switching of the states and the entries in the associated state and error covariance matrices as explained in Algorithm 1. We use the following continuous-discrete system model (Eqn. 7) derived from the dynamics model (Eqn. 1) for human walking with the states switching roles (ϕ_1, $\dot{\phi}_1$ becomes ϕ_2, $\dot{\phi}_2$ respectively and vice versa) when collision occurs:

$$\begin{gathered}\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\\ \dot{x}_3(t)\\ \dot{x}_4(t)\end{bmatrix} = \begin{bmatrix}\dot{\phi}_1(t)\\ \dot{\phi}_2(t)\\ f_1(t,\phi_1(t),\phi_2(t),\dot{\phi}_1(t),\dot{\phi}_2(t))\\ f_2(t,\phi_1(t),\phi_2(t),\dot{\phi}_1(t),\dot{\phi}_2(t))\end{bmatrix} + \begin{bmatrix}0\\0\\w_1(t)\\w_2(t)\end{bmatrix},\\ w(t)\sim\mathcal{N}(0,Q), w(t)=[w_1(t)\ w_2(t)]^T\\ \begin{bmatrix}y_1(k)\\y_2(k)\end{bmatrix} = \begin{bmatrix}\phi_1(k)\\\phi_2(k)\end{bmatrix} + \begin{bmatrix}v_1(k)\\v_2(k)\end{bmatrix}, v(k)\sim\mathcal{N}(0,R)\\ \mathcal{E}(w(t_1)w(t_2)^T)=\delta(t_1-t_2)Q,\ \mathcal{E}(v(i)v(j)^T)=\delta_{ij}R,\\ \mathcal{E}(v(k)w(t)^T)=\mathbf{0},\ \forall i,j, v(k)=[v_1(k)\ v_2(k)]^T\end{gathered} \tag{7}$$

Algorithm 1 Pseudocode for UKF implemented for a continuous-discrete system model.

Step 1. Initialization

$$\underline{\hat{x}}_0 = [\phi_1,\phi_2,\dot{\phi}_1,\dot{\phi}_2]^T_{optimized}, P_{x0} = diag(0.01,0.01,0.5,0.5)$$
$$Q_0 = diag(10,10), R_0 = diag(0.01,0.01)$$
$$P_0^a = \begin{bmatrix}P_{x_k} & 0 & 0\\ 0 & Q_k & 0\\ 0 & 0 & R_k\end{bmatrix}$$

Step 2. For k=1,...,N

Step 2.1 Calculate sigma-points $\{\chi^a_{k,i}\in\mathcal{R}^L | i=0,...,2L\}$, where $L = n_x(=4)+n_w(=2)+n_v(=2)$:

$$\underline{\chi}^a_{k-1} = \begin{bmatrix}\underline{\hat{x}}^a_{k-1} & \underline{\hat{x}}^a_{k-1}+\gamma S^a_{x_{k-1}} & \underline{\hat{x}}^a_{k-1}-\gamma S^a_{x_{k-1}}\end{bmatrix},$$
$$P^a_{k-1} = S^a_{x_{k-1}}(S^a_{x_{k-1}})^T$$

Step 2.2 Time update equations for $i=0,...,2L$:

$$\underline{\chi}_{t|k-1,i} = \mathbf{f}\left(t,\underline{\chi}^x_{k-1,i},\underline{\chi}^w_{k-1,i}\right), t\in[0,\Delta t]$$

Step 2.3 If collision occurred at $t=t_c$

$$\dot{\phi}^-(t_c) = [\chi_{t_c,i,3}\ \chi_{t_c,i,4}]^T$$
$$\dot{\phi}^+(t_c) = \left(\mathbf{T}^{+T}M\mathbf{T}^{+T}\right)^{-1}\mathbf{T}^{+T}\left(S+M\mathbf{T}^-\dot{\phi}^-(t_c)\right)$$

Swap the states (stance leg becomes swing leg and vice versa)

$$\underline{\chi}_{t_c|k-1,i} = [\chi_{t_c,i,2}\ \chi_{t_c,i,1}\ \dot{\phi}^+_2(t_c)\ \dot{\phi}^+_1(t_c)]^T$$
$$\underline{\chi}_{t|t_c,i} = \mathbf{f}\left(t,\underline{\chi}^x_{t_c|k-1,i},\underline{\chi}^w_{k-1,i}\right), t\in[t_c,\Delta t]$$
$$\underline{\chi}_{k|k-1,i} = [\chi_{\Delta t,i,2}\ \chi_{\Delta t,i,1}\ \chi_{\Delta t,i,4}\ \chi_{\Delta t,i,3}]^T$$

Step 2.4 If no collision occurred

$$\underline{\chi}_{k|k-1,i} = [\chi_{\Delta t,i,2}\ \chi_{\Delta t,i,1}\ \chi_{\Delta t,i,4}\ \chi_{\Delta t,i,3}]^T$$
$$\underline{\hat{x}}^-_k = \sum_{i=0}^{2L} w^m_i\underline{\chi}^x_{k|k-1,i}$$
$$P^-_{x_k} = \sum_{i=0}^{2L} w^c_i\left(\underline{\chi}^x_{k|k-1,i}-\underline{\hat{x}}^-_k\right)\left(\underline{\chi}^x_{k|k-1,i}-\underline{\hat{x}}^-_k\right)^T$$

Step 2.5 Measurement-update equations for $i=0,...,2L$:

$$\underline{\chi}^y_{k|k-1,i} = \mathbf{h}\left(\chi^x_{k-1,i},\chi^w_{k-1,i}\right)$$
$$\underline{\hat{y}}^-_k = \sum_{i=0}^{2L} w^m_i\underline{\chi}^y_{k|k-1,i}$$
$$P_{eyey} = \sum_{i=0}^{2L} w^c_i\left(\underline{\chi}^y_{k|k-1,i}-\underline{\hat{y}}^-_k\right)\left(\underline{\chi}^y_{k|k-1,i}-\underline{\hat{y}}^-_k\right)^T$$
$$P_{exey} = \sum_{i=0}^{2L} w^c_i\left(\underline{\chi}^x_{k|k-1,i}-\underline{\hat{x}}^-_k\right)\left(\underline{\chi}^y_{k|k-1,i}-\underline{\hat{y}}^-_k\right)^T$$

Step 2.6 If collision occurred swap the entries in the state covariance matrices $P^-_{x_k}$, P_{eyey}, and P_{exey} to accommodate for the swapping of the states.

Step 2.7 Kalman update equations

$$K_k = P_{exey}P^{-1}_{eyey}$$
$$\underline{\hat{x}}_k = \underline{\hat{x}}^-_k + K_k\left(y_k-y^-_k\right)$$
$$P_{x_k} = P^-_{x_k} - K_kP^{-1}_{eyey}K^T_k$$

End for

where $\{w^i\}$ is a set of scalar weights given by:

$$w^m_0 = \frac{\lambda}{L+\lambda}, w^c_0 = \frac{\lambda}{L+\lambda}+(1-\alpha^2+\beta),$$

$$w_i^m = w_i^c = \frac{1}{2(L+\lambda)}, i = 1, ..., 2L$$
$$\lambda = \alpha^2(L+\kappa) - L, \gamma = \sqrt{L+\lambda}$$

It was observed that arbitrarily choosing the process and measurement covariances results in the state covariance matrix being negative definite due to round-off errors during the filtering process, in which case the filtering cannot continue. We provide more weight to the measurements and so, choose large values for the process covariance.

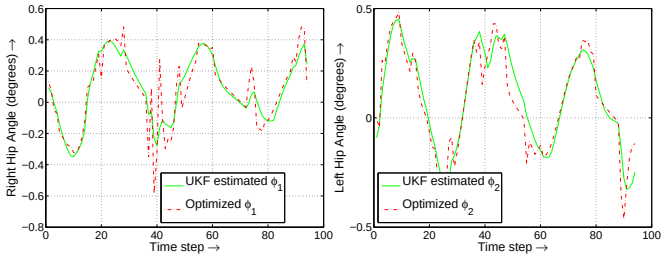
VII. Results

Fig. 5 shows the filtered pose estimates for the hip angles obtained using the UKF. As is evident, the higher frequency noise in the raw pose estimates is filtered and the filtered estimates are temporally coherent. Table 7 shows the raw estimates corresponding to right/left leg forward, simulated, optimized and filtered estimates for the hip angles rendered on the original video. Fig.s. 7(a) and 7(b) show the raw estimates corresponding to the frames with missing pose estimates rendered with reduced color intensity. Please note that the limb identities are not maintained (left leg becomes right and vice versa) in these raw estimates. Fig. 7(c) shows that the hip angles generated using the estimated dynamics very closely explain the true hip angles, even for the frames where no raw pose estimates were available. Note that a subset of pose variables (left/right knee angles, torso location, and heading estimate) for the frames with missing pose estimates is obtained via linear interpolation.

Fig. 7(d) shows the optimized pose estimates obtained by choosing the correct raw pose estimates, where they were available, and substituting the dynamics based pose estimates, in case the raw estimates were missing. Fig. 7(e) shows the pose estimates obtained after filtering the optimized pose estimates using an UKF with the estimated dynamics model as the process model. The filtered pose clearly shows that not only the poor raw pose estimates were replaced by better pose estimates, the frames for which no pose estimates were present were filled with very accurate pose estimates.

Table II presents the results obtained after filtering the optimized pose estimates for a video in the DARPA corpus. Note that the error metric [23] for filtered pose should ideally not be compared with that of the raw estimates. This is because the frames where no pose estimates were available is simply filled with the dynamics estimates and interpolation i.e. no evidence from the images is used. Nevertheless, the error metric considering the total number of frames is better than the raw estimates. Furthermore, the filtered pose estimates are now available for 94 frames as opposed to the 52 frames with noisy raw estimates.

Fig. 6 shows the error distribution across the 13 salient marker points considered for error metric evaluation for the same video. There is significant reduction in the error corresponding to the left/right foot markers as there is no swapping of the limbs in the estimated pose. However, the estimates for the markers in general are not comparable as the number of frames considered are different for the two histograms.



(a) **(b)**

Fig. 5: Results for the UKF filtered states for a video in the DARPA corpus. ($\alpha = \sqrt{\frac{2}{L}}$, $\beta = 0$, $\kappa = \frac{L}{2}$) (a) Right hip angle, and (b) Left hip angle.

Output	Average Error Per Marker Per Frame	Number of Frames Processed
Raw	18.33 pixels	52
Filtered	16.54 pixels	94

TABLE II: Error metric (L1 norm for 13 salient marker points in the image frame) for the raw and the filtered pose estimates for a video in the DARPA corpus. The original frame resolution is 1280×720 and the ground truth was obtained using manual annotation.

VIII. Conclusion

In this work, we propose a technique for estimating the dynamics of a human, solely based on its uncalibrated monocular video. An UKF is used to filter out the optimized gap-filled pose estimates using the continuous-discrete state-space model of the estimated dynamically-equivalent human dynamics model. Results showed that the framework not only resulted in more accurate pose estimates, but also provided physically plausible pose estimates for frames where the original human pose estimation framework failed to provide one. In the future, we plan to build multi-model adaptive Unscented Kalman filter with more complex dynamics model (also simulating knee flexion, torso lean, and arms) that can cater multiple human activities.

Acknowledgments

The authors gratefully acknowledge the support from Defense Advanced Research Projects Agency Mind's Eye Program (W911NF-10-2-0062). The keen insight of the anonymous reviewers is also gratefully acknowledged.

References

[1] P. Agarwal, S. Kumar, J. Ryde, J. Corso, and V. Krovi. An Optimization Based Framework for Human Pose Estimation in Monocular Videos. In *International Symposium on Visual Computing*, 2012.

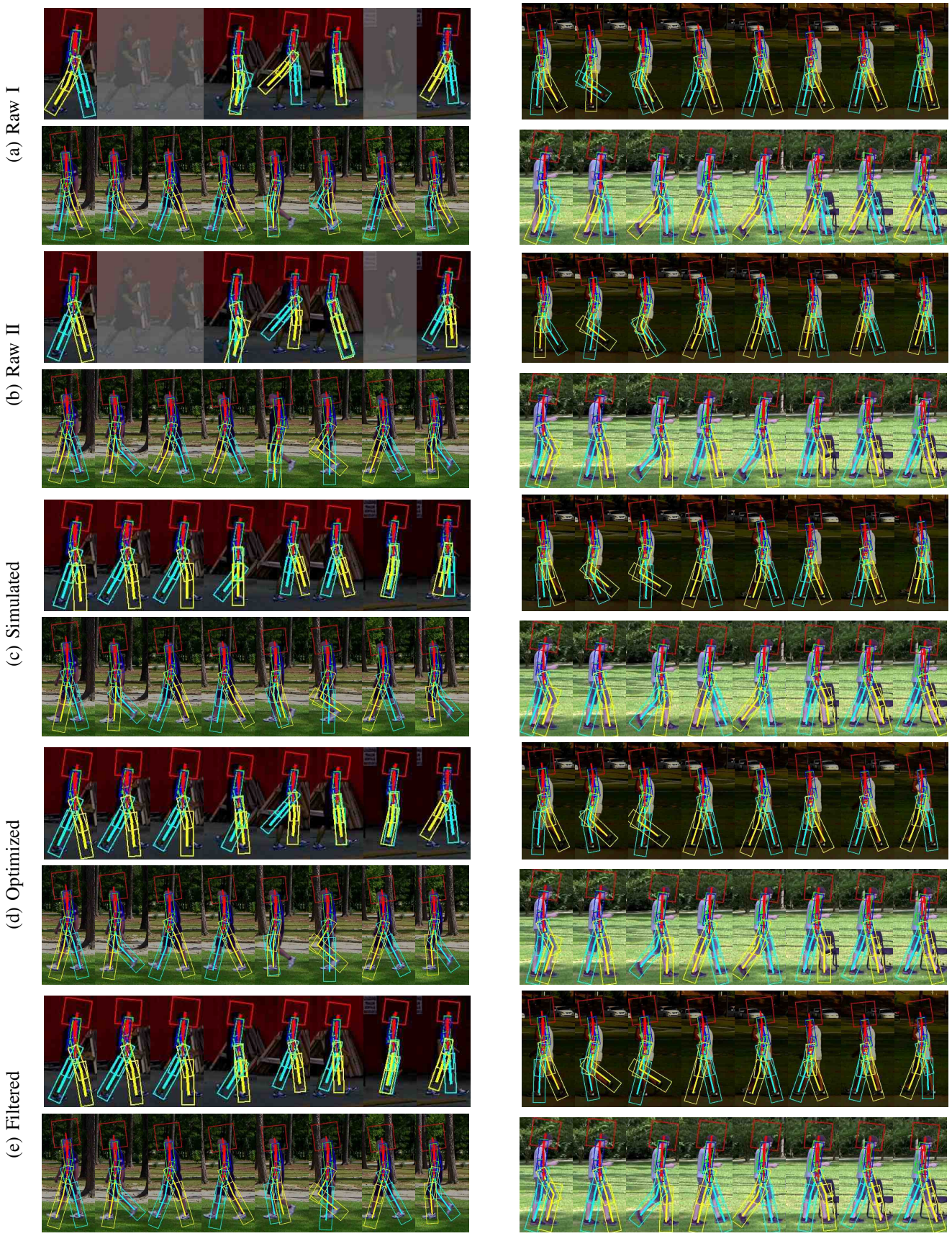
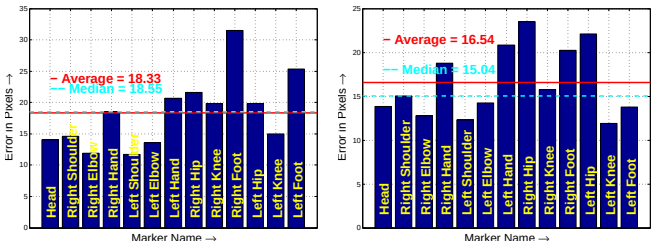


Fig. 7: Human pose estimation results at various stages for videos in the DARPA corpus. (a) Raw estimates with left leg forward, (b) Raw estimates with right leg forward, (c) Optimized simulated dynamics, (d) Optimized estimates, and (e) Filtered estimates. (Please view in color)



(a) **(b)**

Fig. 6: Average error distribution for 13 markers for the obtained pose estimates for a video in the DARPA corpus. (a) Raw estimates averaged for both the outputs (with left leg forward, with right leg forward), (b) Filtered estimates.

[2] A.O. Balan and M.J. Black. An adaptive appearance model approach for model-based articulated object tracking. In *CVPR*, volume 1, pages 758–765, 2006.

[3] A.O. Balan, L. Sigal, M.J. Black, J.E. Davis, and H.W. Haussecker. Detailed human shape and pose from images. In *CVPR*, pages 1–8, 2007.

[4] M. A. Brubaker, L. Sigal, and D. J. Fleet. Estimating contact dynamics. In *ICCV*, pages 2389–2396, 2010.

[5] M.A. Brubaker and D.J. Fleet. The kneed walker for human pose tracking. In *CVPR*, pages 1–8, 2008.

[6] M.A. Brubaker, D.J. Fleet, and A. Hertzmann. Physics-based person tracking using simplified lower-body dynamics. In *CVPR*, pages 1–8, 2007.

[7] A. Bruderlin and T.W. Calvert. Goal-directed, dynamic animation of human walking. *ACM SIGGRAPH Computer Graphics*, 23(3):233–242, 1989.

[8] DARPA. "DARPA Mind's Eye Year 1 Videos", December 2011. `www.visint.org`.

[9] A. Elgammal and C.S. Lee. Inferring 3D body pose from silhouettes using activity manifold learning. In *CVPR*, volume 2, pages II–681, 2004.

[10] H. Fujiyoshi and A.J. Lipton. Real-time human motion analysis by image skeletonization. In *IEEE Workshop on Applications of Computer Vision*, pages 15–21, 1998.

[11] D.E. Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-wesley, 1989.

[12] O.K. Gupta and R.A. Jarvis. Robust pose estimation and tracking system for a mobile robot using a panoramic camera. In *IEEE Conference on Robotics Automation and Mechatronics*, pages 533–539, 2010.

[13] Y.W. Hen and R. Paramesran. Single camera 3d human pose estimation: A review of current techniques. In *International Conference for Technical Postgraduates*, pages 1–8, 2009.

[14] L. Herda, R. Urtasun, and P. Fua. Hierarchical implicit surface joint limits for human body tracking. *Computer Vision and Image Understanding*, 99(2):189–209, 2005.

[15] S.J. Julier and J.K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3): 401–422, 2004.

[16] H. Kjellstrom, D. Kragic, and M.J. Black. Tracking people interacting with objects. In *CVPR*, pages 747–754, 2010.

[17] A.D. Kuo. A simple model of bipedal walking predicts the preferred speed–step length relationship. *Journal of Biomechanical Engineering*, 123:264, 2001.

[18] J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147, 1998.

[19] Tad McGeer. Passive Dynamic Walking. *The International Journal of Robotics Research*, 9(2):62–82, 1990.

[20] M.B. Nogueira, A.A.D. Medeiros, P.J. Alsina, and N.R.N. Brazil. Pose Estimation of a Humanoid Robot Using Images From a Mobile External Camera. In *IFAC Workshop on Multivehicle Systems*, 2006.

[21] J.P.B. Rubio, C. Zhou, and F.S. Hernández. Vision-based walking parameter estimation for biped locomotion imitation. *Computational Intelligence and Bioinspired Systems*, pages 677–684, 2005.

[22] M. Sarkis, K. Diepold, and K. Huper. Pose estimation of a moving humanoid using Gauss-Newton optimization on a manifold. In *7th IEEE-RAS International Conference on Humanoid Robots*, pages 228 –234, 2007.

[23] L. Sigal and M.J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 120, 2006.

[24] L. Sigal and M.J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, volume 2, pages 2041–2048, 2006.

[25] L. Sigal, A. Balan, and M.J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. *Advances in Neural Information Processing Systems*, 20:1337–1344, 2007.

[26] L. Sigal, M. Isard, H. Haussecker, and M.J. Black. Loose-limbed People: Estimating 3D Human Pose and Motion Using Non-parametric Belief Propagation. *IJCV*, pages 1–34, 2011.

[27] C. Sminchisescu and A. Jepson. Variational mixture smoothing for non-linear dynamical systems. In *CVPR*, volume 2, pages II–608, 2004.

[28] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *CVPR*, volume 1, pages I–69, 2003.

[29] R. Urtasun, D.J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *ICCV*, volume 1, pages 403–410, 2005.

[30] R. Urtasun, D.J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *CVPR*, volume 1, pages 238–245, 2006.

[31] M. Vondrak, L. Sigal, and O.C. Jenkins. Physical simulation for probabilistic motion tracking. In *CVPR*, pages 1–8, 2008.

[32] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011.