Probabilistic 3D Tracking: Rollator Users' Leg Pose from Coronal Images

Samantha Ng Adel Fakih Adam Fourney Pascal Poupart
John Zelek
University of Waterloo
{sjng, afakih, afourney, ppoupart, jzelek}@uwaterloo.ca

Abstract

Understanding the human gait is an important objective towards improving elderly mobility. In turn, gait analyses largely depend on kinematic and dynamic measurements. While the majority of current markerless vision systems focus on estimating 2D and 3D walking motion in the sagittal plane, we wish to estimate the 3D pose of rollator users' lower limbs from observing image sequences in the coronal (frontal) plane. Our apparatus poses a unique set of challenges: a single monocular view of only the lower limbs and a frontal perspective of the rollator user. Since motion in the coronal plane is relatively subtle, we explore multiple cues within a Bayesian probabilistic framework to formulate a posterior estimate for a given subject's leg limbs. This paper describes four cues based on three features to formulate a pose estimate: image gradients, colour and anthropometric symmetry. Our appearance model is applied within a non-parametric (particle) filtering system to track the lower limbs. Our tracking system does not rely on any detection for automatic initialization. Preliminary experiments are promising, showing that the algorithm may provide an indication of relative depth for each lower limb.

1. Introduction

Rollators (wheeled walkers) can help older adults by increasing their mobility, facilitating exercise and enhancing safety. Current designs help users with balance by providing two additional points of contact for the upper limbs where some load is transferred to the rollator. This enables users with weak or unhealthy legs to walk more easily. Building on a rollator instrumented with various sensors [27] including two cameras at the Toronto Rehabilitation Institute, our research team at the University of Waterloo is working towards the development of smart rollators that can monitor users and assist them with various tasks. One of our short term goals is to estimate and track the pose of a user's lower limbs with a monocular camera mounted on the rollator.

There are two monocular cameras on the rollator: one forward-facing and one rear-facing (where the field of view encompasses a front view of the user's lower limbs and the frame of the rollator). The possibility of extracting 3D pose information from a markerless rear-facing camera system (such as on the instrumented rollator) is particularly attractive for gait analysis because it allows researchers to collect information in a natural environment as opposed to a lab environment. Research into using the rear-facing camera for gait analysis of rollator users is ongoing. In this paper we describe an appearance model for 3D pose estimation that utilizes multiple cues. The model is incorporated into a Bayesian probabilistic tracking system with non-parametric (particle) filtering. We present some preliminary results from two short video sequences in Sections 3 and 4.1.

There are a few groups that are currently or have worked on intelligent walkers, namely groups at Virginia [30], CMU [8], Utah [12] and Japan [9]. However, we are unique in our work in that we do not limit the environment and that we rely heavily on low-cost and low-power visual sensors. The rollator application thus presents a unique set of challenges. First, only the lower limbs of the user are captured by the rear-facing camera. This limits the contextual information with which we could narrow our search for lower-limb pose. Second, the image plane is perpendicular to the planes of greatest motion for lower limbs, making these motions more difficult to observe. Since joint angles are not very salient from the camera's perspective, it becomes yet another challenge to estimate the length of each limb segment to be tracked. However, step width variability is a strong indicator of frontal plane balance control, and has been correlated with frequency of falls in older adults [ref]. With respect to observing step width variability, the front profile provided by the rear-facing camera on the rollator is highly advantageous compared to prevalent work (e.g. [18]) that tracks the lower limbs from a side profile. Finally, the camera is rigidly attached to the rollator frame and therefore the background moves with respect to the camera's reference frame. Thus, subtraction algorithms such as in [10]

978-0-7695-3651-4/09 $25.00 © 2009 IEEE
DOI 10.1109/CRV.2009.45

and [25] for static backgrounds are not applicable here.

2D pose estimation and tracking of human subjects has been extensively explored for both full-body and partial-body models. An overview of this work is given in [14]. If in the future it becomes possible to mount an additional camera pointing to the torso, then 2D tracking methods based on full-body models may become useful for our application. With regard to multiple cameras, 3D limb tracking has also been widely addressed and reliably implemented (e.g. [1]). Hardware portability, limited power supply and space constraints prevent us from installing a stereo vision system on the walker. Stereo vision would likely compliment our tracking system if and when it becomes tractable. To a lesser extent, recent literature has also focused on 3D tracking from monocular sequences (e.g. [20], [22], [28], [29]). However, these approaches often rely on full-body models for contextual cues. For example in the initialization proposed by [16], knowledge of the human torso connecting to the thighs is used to prune possible body configurations. Unfortunately we cannot observe the torso with our apparatus. There has been some research into monocular 3D tracking with partial-body models. In [3], arm limb segments are identified using action templates [5]. These templates however depend on significant motion being almost parallel to the image plane. As well, initialization of physical model parameters such as length of segments depend on being able to observe the joint angles.

To our knowledge, there is no system that addresses all of the constraints of our rollator application. As a first step in addressing our problem, we explore a general appearance model, based on multiple cues, that requires no detection in an initialization phase. The cues that we use are quite modest; image gradients (e.g. [23]), colour (e.g. [13]) and anthropometric symmetry have been exploited in several works. However, these low-level cues along with edges, corner features (ex. [24]), etc., tend to be used for segmentation (e.g. [19]) and body-part detection. Another class of detectors uses exemplar matching (e.g. [26],[15]). In future we may incorporate a detection method into our initialization, but for the present we can demonstrate promising preliminary results with a purely probabilistic framework.

Many authors including [21], [4] address the problem of slow convergence for 3D monocular tracking with a sampling approach and high-dimensional state vector (in this case, 20 state elements as described in Section 2.3). M. Black in [2] addresses convergence issues when particle filtering is applied to tracking motion boundaries, another important cue that we plan to investigate in future. Here, we simply apply the Condensation algorithm proposed in [11], as a preliminary tracking algorithm to qualitatively evaluate the strength of our appearance model, but do note the limitations of the current algorithm for future work.

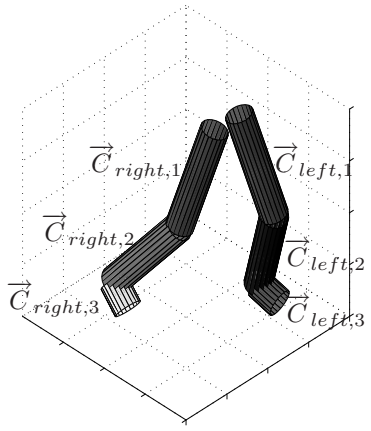


Figure 1. The 3D model.

2. Problem formulation

2.1 Physical model

We adopt a model composed of regular cylinders for each leg segment: thigh, calf and foot. We define the state vector $\overrightarrow{X}$, from which the position and orientation of each cylinder $\overrightarrow{C}_k$ in the model in Figure 1 can be determined. There are 20 elements in the state vector $\overrightarrow{X}$: the position of the right hip, the spherical coordinates of the left hip relative to the right hip, the lengths of the cylinders (assuming symmetry between left and right legs), a single width for all cylinders, 3 DOF joint angles for the hips, 1 DOF joint angles for the knees, and 1 DOF joint angles for the ankles.

Anthropometric constraints are enforced on the proportional lengths and widths of each limb segment according to tables in [31]. Further, we constrain the ranges of absolute lengths, widths and joint angles according to 5th and 95th percentile statistics in [17]. Finally, we enforced a constraint that there must always be at least one foot on the ground (where the location and orientation of the ground plane relative to the camera was physically measured).

2.2 Model projection to the image plane

For each pixel location ij in the image plane, we define the function $s(i, j, \overrightarrow{X})$:

$$s(i,j,\overrightarrow{C}_{left,k}) \stackrel{\text{def}}{=} \begin{array}{ll} 1 & \text{if } i,j \in \Pi(\overrightarrow{C}_{left,k}) \\ 0 & \text{otherwise} \end{array}$$

$$s(i,j,\overrightarrow{C}_{right,k}) \stackrel{\text{def}}{=} \begin{array}{ll} 1 & \text{if } i,j \in \Pi(\overrightarrow{C}_{right,k}) \\ 0 & \text{otherwise} \end{array}$$

$$s(i,j,\overrightarrow{C}_k)=\min\Big(s(i,j,\overrightarrow{C}_{left,k})+s(i,j,\overrightarrow{C}_{right,k}),1\Big) \quad (1)$$

$$s(i,j,\overrightarrow{X})=\min\Big(\sum_k s(i,j,\overrightarrow{C}_k),1\Big) \quad (2)$$

where $\Pi(\overrightarrow{X})$ is the projection of the model on the image plane:

$$\Pi([x_1,x_2,x_3]^T)=\Big[\frac{x_1}{x_3},\frac{x_2}{x_3}\Big]^T \quad (3)$$

2.3 Formulation

We formulate the estimation problem as a dynamic system:

$$\begin{cases} \text{State:} & \overrightarrow{X}(t+1) = f(\overrightarrow{X}(t)) + \overrightarrow{n}_s(t), \\ & \overrightarrow{n}_s(t) \sim \mathcal{N}(0,\Sigma_s), \\ \text{Measurement:} & I_c(t) = g(\overrightarrow{X}(t)) + \overrightarrow{n}_m(t), \\ & \overrightarrow{n}_m(t) \sim \mathcal{N}(0,\Sigma_m) \end{cases} \quad (4)$$

where $I(t)$ is the observed image at time t and $I_c(t)$ is a set of image cues c extracted at t.

Our aim is to determine at every time instant t, the probability distribution $P(\overrightarrow{X}(t)|t)$ of the state-vector given the image measurements from 0 to t.

The state equation provides a means to predict $P(\overrightarrow{X}(t+1)|t)$ from $P(\overrightarrow{X}(t)|t)$. From the measurement equation, the likelihood of the state vector given the image measurement $P(I(t+1)|\overrightarrow{X}(t+1))$ at $t+1$ can be determined. Bayes rule permits us then to infer the posterior probability:

$$P(\overrightarrow{X}(t+1)|t+1) = \frac{P(\overrightarrow{X}(t+1)|t)P(I(t+1)|\overrightarrow{X}(t+1))}{\int P(\overrightarrow{X}(t+1)|t)P(I(t+1)|\overrightarrow{X}(t+1))d\overrightarrow{X}} \quad (5)$$

3. Image appearance and likelihood

Four image cues are used to determine the likelihood of the state vector given the image. Since these cues are used within a probabilistic framework, they do not need to be perfect, but somewhat indicative of the pose. In Section 4, these cues are combined in a particle filtering model for state tracking.

3.1 Image gradients

If we assume homogeneity within leg regions, then these regions are characterized by a very low average gradient magnitude. The top row of images in Figure 2 shows three rollator users, each with different clothing. The middle row shows corresponding gradient maps and the final row plots the pixel-column sum of gradients across each image, normalized to the ranges of gradient magnitude within each image. Figure 2(a) illustrates the horizontal position of the legs clearly noticeable from the image's gradients. In Figure 2(b), the leg regions are again aligned with regions of low gradients. However, there are limitations to using gradients as a cue. An *extra* leg not belonging to the walker is also implied at the right of Figure 2(b) by the regions of low gradients. Thus, gradients do not discriminate between legs of different people. Further, the background on the left side of the scene is fairly uncluttered, which could lead to *false positives*. Figure 2(c) illustrates opposite drawbacks; the background scene is cluttered but the leg regions also contain gradients from folds in clothing. However the leg positions are still noticeable from the gradient plot.

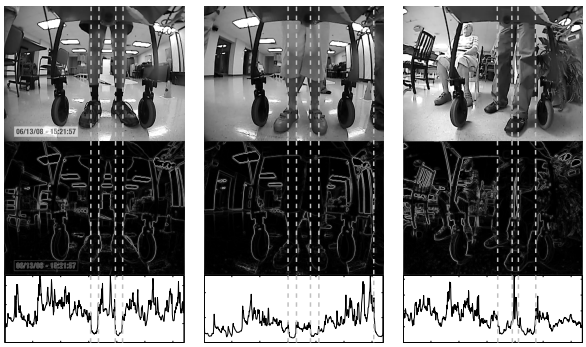


Figure 2. Image gradient magnitudes indicating the position of the lower limbs.

To measure the observation, a given grayscale image $I_{grayscale}$ is first smoothed with a Gaussian kernel. The resulting image is then convolved with two 3×3 Sobel kernels, horizontal and vertical, to produce a gradient magnitude image $G(I_{grayscale})$. $G(I_{grayscale})$ is shifted so that $min_{i,j}\,(G(I_{grayscale},i,j)) = 0$ and $\sum_{i,j} G(I_{grayscale},i,j) = 1$. Given the gradient observation, we assign a likelihood to a state hypothesis $\overrightarrow{x}^{[n]}$:

$$P(I_{gradient}|\overrightarrow{x}^{[n]})=\lambda_{grad}\exp\Big(-\lambda_{grad}\frac{\sum_{i,j} s(i,j,\overrightarrow{x}^{[n]})G(I_{gradient},i,j)}{\sum_{i,j} s(i,j,\overrightarrow{x}^{[n]})}\Big) \quad (6)$$

where λ_{grad} influences the spread of the exponential distribution.

3.2 Colour

The observation of uniformity within leg regions can be applied in colour space. We transform the observed image I to the normalized RGB colour space since it has been shown to be robust to illumination changes and folds in

clothing [6] and has lower dimensionality than for example *RGB colour space*. An evaluation of alternate colour spaces, not included here, is certainly an important area for further exploration. From I_{colour}, three pairs 3D histograms are constructed, one for each pair of leg segments. Figure 3 illustrates homogeneity of colour in each pair of leg segments. We consider pairs of segments rather than considering the whole leg as one region, in order to encode spatial information. For example, in Figure 3, the thigh and calf segments have different histograms than the foot segments.

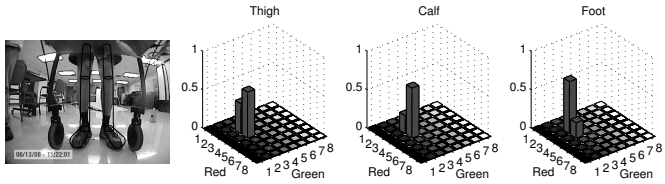


Figure 3. Histogram of normalized RG content in leg segment pairs.

For each pixel $I_{clr}(i,j)$, we exclude its contribution to the histogram of its corresponding segment pair, normalize the resulting histogram and evaluate it at $I_{clr}(i,j)$, giving $H_{clr}(i,j)$. If there are fewer, stronger modes in the histogram (the colour is more uniform) then $H_{clr}(i,j)$ tends to be large for a majority of i,j. The likelihood of a given state hypothesis is computed by:

$$P(I_{clr}|\overrightarrow{x}^{[n]})=\lambda_{clr}\exp\left(-\lambda_{clr}\frac{\sum_{i,j,k} M(i,j)s(i,j,\overrightarrow{C}_k^{[n]})H_{clr}(i,j)}{\sum_{i,j,k} M(i,j)s(i,j,\overrightarrow{C}_k^{[n]})}\right) \quad (7)$$

where λ_{clr} controls the spread of the distribution and $M(i,j)$ is a subsampling mask.

3.3 Symmetry between left and right segments

Here we exploit the observation that people tend to exhibit symmetry in their left and right body segments. Preliminary experiments indicated that gray-value histograms actually performed better than normalized RGB colour histograms, providing a stronger signal for comparing two given segments. We therefore compute a gray-level normalized histogram $H_{sym}(\Pi(\overrightarrow{C}_{left,k}))$ for a left segment k and compare it with its right counterpart $H_{sym}(\Pi(\overrightarrow{C}_{right,k}))$ by straightforward differencing:

$$\Delta H=\sum_m |H_a(m)-H_b(m)| \quad (8)$$

with m a histogram bin. The likelihood based on symmetry is then determined as follows:

Figure 4. Extracting local windows around a foot projection.

$$P(I_{sym}|\overrightarrow{x}^{[n]})=\lambda_{sym}\eta_{sym}\exp\left(-\lambda_{sym}\sum_k \Delta H_{sym}^{[n]}(k)\right) \quad (9)$$

where λ_{sym} controls the spread of the distribution. η_{sym} is simply a normalizing constant to account for $0 \leq \Delta H_{sym}(k) \leq 2$.

3.4 Contrast between the foot segments and the floor

The areas immediately in front of the rear wheels tend not to include clutter. Since the location of the rollator frame is fixed relative to the camera, we therefore know the location of the rollator wheels in the image at all times. We create a gray-level histogram of the floor H_{floor} from a local window of pixels known to be immediately infront of the image of the wheels. We compare gray-level normalized histograms of local windows around the boundaries of the feet projections (and not intersecting the calf projections) to the histogram of the floor, following the same method of histogram comparison as in Equation 8. Figure 4 illustrates finding local circular windows around the boundary of the right foot projection. Rays are extended from the centroid of the projection outward at equal angular intervals. The intersection of these rays and the projection boundary are the centers of the windows. We use a window radius equal to half the projected width of the foot segment.

The pixels within a given window are divided into two sets, foreground and background. Pixel i,j is in the foreground set when $s(i,j,\overrightarrow{C}_{k=3}) = 1$ and the background set otherwise. Gray-level normalized histograms are computed for each set and compared to H_{floor} using Equation 8. We then have for foreground and background sets of a given window, $\Delta H_{fg,floor}$ and $\Delta H_{bg,floor}$ respectively. If the projection of the feet are accurate then we would expect a large $\Delta H_{fg,floor}$ relative to $\Delta H_{bg,floor}$

We calculate a likelihood given the observation of symmetry as follows:

$$P(I_{floor}|\overrightarrow{x}^{[n]})=\lambda_{floor}\exp\left(-\lambda_{floor}\frac{1}{L}\sum_{l=1}^{L}\frac{\Delta H_{fg,floor}^{[n]}(l)}{\Delta H_{bg,floor}^{[n]}(l)}\right) \quad (10)$$

where λ_{floor} controls the spread of the distribution and L is the number of circular windows being evaluated.

3.5 Observation cue evaluation

The likelihoods for a set of N state hypotheses ($n = 1...N$) given observation cue c is compared to a pseudo-ground-truth weight distribution w_{true} of the same set of particles as follows. For each cylinder k in the true image foreground $s(i,j,\overrightarrow{\mu}_x)$, a centroid coordinate $C_{true}^{[k]}$, major orientation $O_{true}^{[k]}$, and pixel area $A_{true}^{[k]}$ is calculated. These three attributes are also calculated for each set of cylinder projections of the state hypotheses. Each hypothesis projection $s(i,j,\overrightarrow{x}^{[n]})$ is compared to $s(i,j,\overrightarrow{\mu}_x)$ using Equation 11, where $w_{true}^{[n]}$ can be considered a pseudo-ground-truth weighting.

$$w_{C,true}^{[n]}=\eta_C\exp\left(\sum_{k=1}^{6}\left(C_{true}^{[k]}-C^{[k][n]}\right)^2\right)$$

$$w_{O,true}^{[n]}=\eta_O\exp\left(\sum_{k=1}^{6}\rho_{true}^{[k]}\left(O_{true}^{[k]}-O^{[k][n]}\right)^2\right)$$

$$w_{A,true}^{[n]}=\eta_A\exp\left(\sum_{k=1}^{6}\left|A_{true}^{[k]}-A^{[k][n]}\right|\right)$$

$$w_{true}^{[n]}=\eta\sqrt{\zeta_C w_{C,true}^{[n]}+\zeta_O w_{O,true}^{[n]}+\zeta_A w_{A,true}^{[n]}} \quad (11)$$

where $\rho_{true}^{[k]}$ is the projected length of cylinder k, and $\zeta_C,\zeta_O,\zeta_A \in [0,\infty)$ scales the contributions of $w_{C,true}$, $w_{O,true}$, and $w_{A,true}$ respectively to w_{true}. $\rho^{[k]}$ scales the contribution of each cylinder's projected orientation. More noise is expected from orientations of shorter limbs, thus $\rho^{[k]}$ is directly proportional to the length of segment k. η is a normalizing constant over all $w_{true}^{[n]}$.

To determine how effective a given cue c is, we can simply measure the percentage of particles $p_{correct}$ whose posterior for that cue, starting from a uniform prior distribution, moved in the correct direction according to Equations 12 and 13.

$$p_{correct}^{[n]} \stackrel{\text{def}}{=} \begin{array}{ll} 1 & \text{if } w_{true}^{[n]} > \frac{1}{N} \text{ and } P(\overrightarrow{x}^{[n]}|I_c) > \frac{1}{N} \\ 1 & \text{if } w_{true}^{[n]} \leq \frac{1}{N} \text{ and } P(\overrightarrow{x}^{[n]}|I_c) \leq \frac{1}{N} \\ 0 & \text{otherwise} \end{array} \quad (12)$$

$$p_{correct}=\frac{1}{N}\sum_n p_{correct}^{[n]} \quad (13)$$

where there are N state hypotheses.

This method, although coarse, avoids being overly sensitive to the somewhat subjective evaluation of w_{true}.

Each appearance cue was evaluated against two rollator users *A* and *B*, shown in Figures 5(a) and 5(b) respectively. In 5(a), the legs are bare and well-separated. In 5(b), the legs are partially covered by wide shorts, a more challenging scenario. True poses were manually segmented for each user, and are also shown in Figures 5(a) and 5(b).

(a) User A (b) User B

Figure 5. Model projection of the true poses for two rollator users.

3000 state hypotheses were randomly generated for each user. The search space was constrained laterally to within the rollator frame, and between 30cm and 130cm depthwise from the camera. For each set of estimates, a pseudo-ground-truth weight distribution was calculated.

Tables 1(a) and 1(b) show the percentage of state hypotheses whose likelihoods given each observation cue were consistent (according to Equation 13) with $w_{true,A}$ and $w_{true,B}$ respectively. For each user, the 3000 hypotheses were sorted according to $w_{true,user}$. The weights of the n' best of these estimates were re-normalized as a distribution over n' and each cue evaluated again for $n' = 1500$, $n' = 750$ and $n' = 375$. This was done in order to observe the cues' performance for different widths of hypotheses distributions.

(a) User A

Cue	Best 375	Best 750	Best 1500	3000
Gradients	0.82	0.83	0.86	0.86
Colour	0.80	0.84	0.84	0.76
Symmetry	0.90	0.88	0.89	0.86
Floor contrast	0.86	0.85	0.84	0.81

(b) User B

Cue	Best 375	Best 750	Best 1500	3000
Gradients	0.79	0.81	0.82	0.84
Colour	0.81	0.81	0.78	0.70
Symmetry	0.82	0.83	0.81	0.78
Floor contrast	0.86	0.85	0.82	0.77

Table 1. Percentage $p_{correct}$ of Hypotheses Likelihoods Consistent with w_{true}

Colour appeared to be the weakest cue, performing particularly poorly against the image of user *B* for large de-

viations in the hypotheses. The background in many areas of the image did not exhibit great enough contrast in normalized colour, resulting in false-positives. The colour cue showed better performance for narrower hypotheses deviations, due to the red component in the users' skincolours contrasting with the background. User *A*'s socks did cause the colour cue to favour hypotheses where the calf segments ended at the top of the socks. This result was not entirely unexpected. It was hoped that the likelihood given the feet-to-floor contrast cue might compensate for the bias. However, as is shown in Section 4.1, the effects of the bias were evident within less than one second of a tracking sequence. The other three cues appeared to perform well and fairly consistently. Symmetry and feet-to-floor contrast tended to increase performance with smaller deviations in the hypotheses, whereas gradient performance decreased slightly with smaller deviations in the hypotheses.

3.6 Combining weights

For simplicity we computing an overall likelihood of the state vector given the image cues as a product of the likelihoods given each cue:

$$P(I|\overrightarrow{x}^{[n]})=\eta_w \prod_{allcues} P(I_{cue}|\overrightarrow{x}^{[n]}) \quad (14)$$

where η_w is a normalizing constant.

The cues are not actully independent of one another given a hypothesis. Currently we are exploring statistical dependencies between image gradients and colour using methods described in [32]. There are more established and sophisticated methods of combining cues such as AdaBoost [7] and its many variants. These methods will be explored when a large enough annotated training set is available.

4 Tracking

We chose a constant-velocity model for prediction:

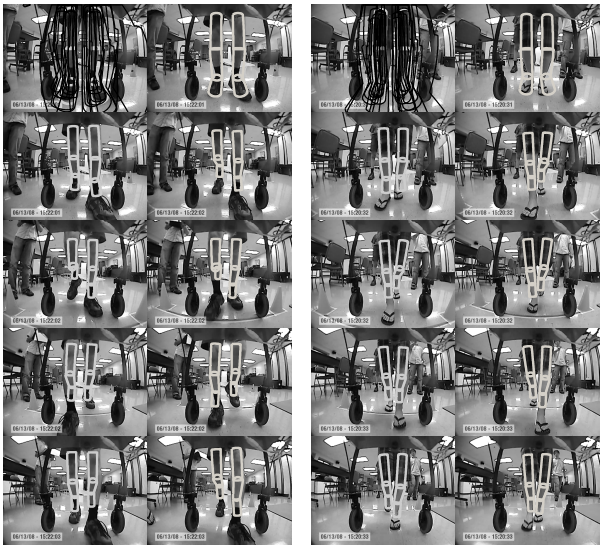
$$f(\overrightarrow{X}(t+1))=\overrightarrow{X}(t)+\frac{d\overrightarrow{X}(t)}{dt}\Delta t \quad (15)$$

A constant-velocity model was chosen because the motion of an elderly rollator user is typically slow and gradual. Referring to Equation 4, the noise parameter $\overrightarrow{n}_s(t)$ is manually estimated. Standard deviation in joint angles are set to 5 degrees, while deviation in limb segment sizes is set to 2cm. Deviation in position is set to 10cm. Estimates violating anatomical constraints are resampled.

The state at $t = 0$ is initialized in the same way as described in Section 3.5, with the additional assumption that the user is starting from a standing position and facing the camera. We apply the Condensation algorithm [11] to approximate the posterior probability given in Equation 5.

4.1 Preliminary evaluation

We apply our appearance model and tracking framework to a 6.3-second video sequence featuring user *A*, and to a 3.9-second video sequence featuring user *B*. The sequences are each subsampled at 10fps. Figures 6(a) and 6(b) show the means of the posterior distributions for users *A* and *B* respectively. For brevity, the results are shown here at 5fps. 10 randomly selected hypotheses are plotted in the initial frames to illustrate that the starting position was not manually initialized to the correct position.



(a) User A (b) User B

Figure 6. From left to right, top to bottom, tracking results for 10 sequential frames at 5fps. Cylinder projections indicate the *mean* of the posterior distribution. Multiple hypotheses are shown in the first frame.

The results do indicate that our appearance model can coarsely infer relative depth. For each frame, the algorithm finds the correct depth of one lower limb relative to its counterpart and we can observe the general walking motion. Further, the projections are laterally (left-right) very well placed. User *B*'s hip angles (about the camera axis) are nicely captured.

The means of the state posterior distributions for each frame are not perfect. Knee and hip joint locations are not estimated properly, but this result is as expected since the cues we have so far implemented do not describe knee characteristics. Also, we do not have observations of the hips; they are hidden by the walker frame. The appearance model favours user *A*'s socks as not being part of the calves. Thus, the user's socks are always interpretted as feet. A similar

problem occured for user *B*. Finally there is also some lag evident in the state predictions, which indicates the need for either a higher derivative form of motion model, or a slightly faster video sampling rate than 10fps.

For the longer video sequence featuring user A, for each frame we compute the average standard deviation in centroidal position of the model cylinders $\overrightarrow{C}_k$ over all hypotheses. Figure 7 shows the standard deviation over 6.3 seconds, separated into three directional components: left-right, craniocaudal (up-down), and dorsoventral (depth).

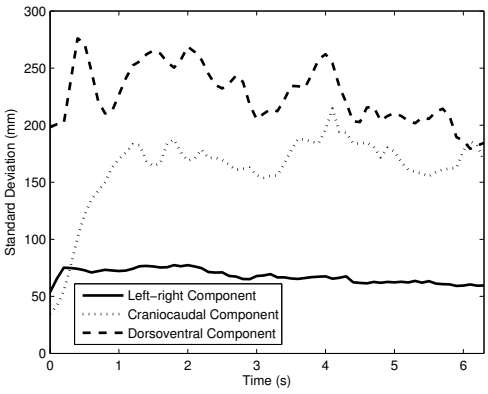


Figure 7. Standard deviation over all hypotheses of centroidal positions of $\overrightarrow{C}_k$, for each frame in the video sequence featuring user A.

In Figure 7, it is important to note that the *initial increase* in standard deviation over the first 5 frames is *due to the fact that we constrain the initial hypotheses to standing positions*. We therefore expect a sharp rise in standard deviation over the first few frames after this constraint is removed.

The left-right component of deviation appears to be stable, averaging approximately 7cm over 6.3 seconds and slowly decreasing. This result supports the qualitative observation that legs are well-tracked in the left-right direction. The craniocaudal component of the standard deviation is greater and less stable, although promisingly it appears not to be pathologically increasing. The dorsoventral component does not seem to be pathologically increasing either, but exhibits even greater, cyclical fluctuations, which agree with the lag qualitatively observed in tracking results for when leg motions change direction. This result is natural, since the majority of limb motion lies in the sagittal plane. Longer video sequences and more iterations of filtering will allow us to more closely study convergence.

5 Conclusions and future work

We present four cues for 3D monocular tracking of rollator users' lower limbs from a coronal perspective. These cues are: homogeneity within leg regions indicated by a low gradient content, uniformity of colour, anthropometric symmetry, and contrast between the gray-level distributions of the floor and the feet. Each cue is evaluated separately against two images of different rollator users and for each user a set of 3000 hypoetheses distributed within the operating space of the rollator frame. Between 70% to 84% of good and bad hypotheses were promoted and demoted respectively. Colour uniformity appears the weakest-performing cue, although its performance increases to par with the other cues when evaluating subsets of hypotheses more tightly distributed about the true poses.

Preliminary tracking results are promising in that the algorithm can capture the continuous alternation of one leg in front of the other over at least 6.3 seconds. The feet-to-floor contrast cue does not contribute strongly enough to the posterior estimation, as the tracking algorithm tends not to demote hypotheses that place the feet where the calves should be. The left-right position component of the lower-limbs are particularly well tracked. From an analysis of the longer of two video sequences used in the tracking evaluation, the average of the standard deviations in centroidal positions of each segment of the lower limbs appears in the left-right component to be stable at approximately 7cm and slowly decreasing. The dorsoventral and craniocaudal components of the deviation are greater (as much as 27cm and 19cm respectively) and exhibit higher fluctuation, which agrees with the lag observed in tracking results for when leg motions change direction. This result is expected, since the majority of limb motion lies in the sagittal plane. Neither the dorsoventral nor craniocaudal components of the deviation show divergent behaviour. A study of convergence over longer video sequences is in progress.

We continue to expand our data sets in order to refine our appearance model and to quantitatively evaluate tracking over longer sequences. There also remains potential for further exploration of salient cues such as discontinuities in optical flow [2]. We also intend to explore methods for handling occlusion and estimating joint position.

6 Acknowledgements

We thank William McIlroy and James Tung from the Toronto Rehabilitation Institute who built and designed the iWalker used to record the videos. This research was funded by a CIHR grant #MIA-85860 for Mobility in Aging and a grant from the UW-Schelegel Research Institute in Aging.

References

[1] O. Bernier, P. Cheung-Mon-Chan, and A. Bouguet. Fast nonparametric belief propagation for real-time stereo articulated body tracking. *Comput. Vis. Image Underst.*, 113(1):29–47, 2009.

[2] M. J. Black and D. J. Fleet. Probabilistic detection and tracking of motion boundaries. *Int. J. Comput. Vision*, 38(3):231–245, 2000.

[3] D. Bullock and J. Zelek. Towards real-time 3-d monocular visual tracking of human limbs in unconstrained environments. *Real-Time Imaging*, 99(7):323–353, November 2005.

[4] K. Choo and D. Fleet. People tracking using hybrid monte carlo filtering. *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, 2:321–328 vol.2, 2001.

[5] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 928– 934, 1997.

[6] P. Fieguth and D. Terzopoulos. Color based tracking of heads and other mobile objects at video frame rates. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 21–27, 1997.

[7] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[8] J. Glover, D. Holstius, M. K. Montgomery, A. Powers, J. Wu, S. Kiesler, J. Matthews, and S. Thrun. A roboticallyaugmented walker for older adults. Technical Report CMU-CS-03-170, Carnegie Mellon University, School of Computer Science, 2003.

[9] Y. Hirata, A. Muraki, and K. Kosuge. Motion control of intelligent walker based on renew of estimation parameters for user state. *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 1050–1055, Oct. 2006.

[10] T. Horprasert, D. Harwood, and L. Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *IEEE ICCV'99 Frame-Rate Workshop*, 1999.

[11] M. Isard and A. Blake. Conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):528, 1998.

[12] V. Kulyukin, A. Kutiyanawala, E. LoPresti, J. Matthews, and R. Simpson. iwalker: Toward a rollator-mounted wayfinding system for the elderly. In *RFID, 2008 IEEE International Conference on*, Las Vegas, NV, 2008. IEEE Computer Society.

[13] A. S. Micilotta, E. J. Ong, and R. Bowden. Detection and tracking of humans by probabilistic body part assembly. In *Proc. of British Machine Vision Conference*, 2005.

[14] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104:90126, 2006.

[15] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part III*, pages 666–680, London, UK, 2002. Springer-Verlag.

[16] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages 326–333, 2004.

[17] NASA. Man-system integration standards nasa-std-3000, July 1995.

[18] H. Ning. Kinematics-based tracking of human walking in monocular video sequences. *Image and Vision Computing*, 22(5):429–441, November 2004.

[19] R. Nock and F. Nielsen. Statistical region merging. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1452–1458, 2004.

[20] L. Sigal and M. J. Black. Predicting 3d people from 2d pictures. In *Articulated Motion and Deformable Objects 2006*, pages 185–195, 2006.

[21] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *The International Journal of Robotics R*, 22:371–391, 2003.

[22] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 69–76, June 18-20 2003.

[23] J. Snoek, J. Hoey, L. Stewart, and R. S. Zemel. Automated detection of unusual events on stairs. *Image and Vision Computing*, 27:153–166, 2009.

[24] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:814–827, 2003.

[25] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 1999.

[26] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *Int. J. Comput. Vision*, 48(1):9–19, 2002.

[27] J. Tung, W. Gage, K. Zabjek, D. Brooks, B. Maki, A. Mihalidis, G. Gernie, and W. McIlroy. iwalker: a real world mobility assessment tool. In *CMBE Conference*, 2007.

[28] R. Urtasuna. Monocular 3d tracking of the golf swing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 932– 938, June 20-25 2005.

[29] M. Vondrak, L. Sigal, and O. C. Jenkins. Physical simulation for probabilistic motion tracking. In *Computer Vision and Pattern Recognition (CVPR 2008)*, 2007.

[30] G. Wasson, P. Sheth, C. Huang, and M. Alwan. Aging medicine. In *Eldercare Technology for Clinical Practitioners (M. Alwan and R. Felder, eds.)*, chapter Intelligent Mobility Aids for the Elderly. Humana Press, 2008.

[31] D. A. Winter. *Biomechanics and Motor Control of Human Movement*. Wiley, Toronto, 2 edition, 1990.

[32] C. Zhou and B. Mel. Cue combination and color edge detection in natural scenes. *Journal of Vision*, 8(4):1–25, 2008.