Towards a Mobility Diagnostic Tool: Tracking Rollator Users' Leg Pose With a Monocular Vision System

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

Abstract— **Cognitive assistance of a rollator (wheeled walker) user tends to reduce the attentional capacity of the user and may impact her stability. Hence, it is important to understand and track the pose of rollator users before augmenting a rollator with some form of cognitive assistance. 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. In this work, our focus is on evaluating the appearance model (the cues). Preliminary experiments indicate that texture and colour cues conditioned on the appearance of a rollator user outperform more general cues, at the cost of manually initializing the appearance offline.**

I. INTRODUCTION

Our research team is developing a smart rollator to monitor and assist users with various tasks of daily living. In particular, we are interested in assisting cognitively impaired users with navigation and obstacle avoidance, as well as reminders and prompts. To that effect, rollators provide a convenient platform that can be instrumented with various sensors and actuators that are less constrained and less intrusive than wearable sensors, but still mobile and therefore capable of accompanying the user in most of his/her movements. Many research groups have already developed smart rollators with various combinations of monitoring and assisting functionalities [1] [2] [3] [4]. However, these functionalities are often developed at the expense of the original purpose of the rollator, which is to ensure stability. [5] points out that navigation assistance may de-stabilize the user unless the steering direction is in line with the user's intent. This can be particularly dangerous for users with dementia since balance control is one of the functions that deteriorates with disease progression. Even users with early dementia are known to have a reduced attentional capacity that often leads to increased gait variability when dual tasking such as listening to a prompt while walking [6]. Since falls are the most important cause of injury for older adults, it is critical to understand the pose and stability of users before augmenting rollators with navigation assistance, prompting or any other form of cognitive assistance.

Our instrumented rollator was designed at the Toronto Rehabilitation Institute [7] and includes two monocular cameras: one forward-facing and one rear-facing. The possibility of extracting 3D pose information from a markerless rear-facing camera system is particularly attractive for gait analysis and a future understanding of the impact on stability of various forms of cognitive assistance. In this paper, we describe a multiple cue appearance model for 3D pose estimation to be used in a Bayesian probabilistic tracking system. Our work differs from other intelligent walker projects [1], [2], [3], [4] in that we do not limit the environment and that we rely heavily on low-cost and low-power visual sensors. Challenges include: (1) only the lower limbs of the user are captured by the rear-facing camera; and (2) 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 [8]. 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. [9]) 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, negating the use of simple background subtraction algorithms.

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 [10]. 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. Hardware portability, limited power supply and space constraints prevent us from installing a stereo vision system on the walker.

There has been some research into monocular 3D tracking with partial-body models. In [11], arm limb segments are identified using action templates [12]. 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.

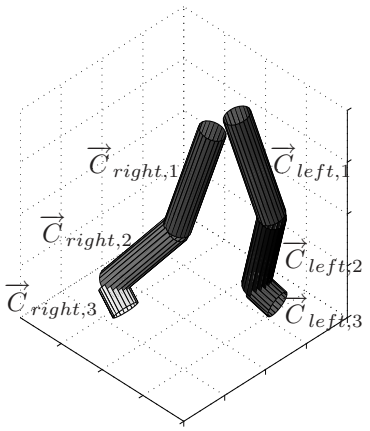


Fig. 1. The 3D model.

To our knowledge, there is no system that adequately addresses all of the constraints of our rollator application. [13] explored a general appearance model, based on multiple cues, that required no detection in an initialization phase. Here we explore the advantages of using a more specific appearance model based on texture and colour, which minimizes our reliance on heuristics.

II. PROBLEM FORMULATION

A. Physical model

We adopt a model composed of tapered 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 19 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), 3 DOF joint angles for the hips, 1 DOF joint angles for the knees, and 1 DOF joint angles for the ankles. The limb segment widths at the hip, knee and ankle (in proportion to the segment lengths) are determined from an initial foreground segmentation (which will be explained in Section III).

Anthropometric constraints are enforced on the proportional lengths and widths of each limb segment according to tables in [14]. Further, we constrain the ranges of absolute lengths, widths and joint angles according to 5th and 95th percentile statistics in [15]. 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).

B. Model projection to the image plane

For each pixel location ij in the image plane, we define the function $s(i,j,\overrightarrow{X})$ equal to 1 if it lies within the projection $\Pi(\overrightarrow{X})$ of the model on the image plane, and 0 otherwise. Similarly, $s(i,j,\overrightarrow{C}_{right,k})$ considers the projection $\Pi(\overrightarrow{C}_{right,k})$ of the right k^{th} segment on the image plane and so on.

C. Formulation

We formulate the estimation problem as a dynamic system:

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

where $I(t)$ is the observed image at time t and $I_c(t)$ is a set of image cues $c = cue1, cue2...$ 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:

$$P(\overrightarrow{X}(t+1)|t+1) = \frac{P(\overrightarrow{X}(t+1)|t)P(I(t+1)|\overrightarrow{X}(t+1))}{\int P(\overrightarrow{X}(t+1)|t)P(I(t+1)|\overrightarrow{X}(t+1))d\overrightarrow{X}} \tag{2}$$

III. IMAGE APPEARANCE AND LIKELIHOOD

Two image cues are used to determine the likelihood of the state vector given the image: colour and texture. In this work, we manually position the 3D model projection to coincide with the image foreground. However, we can assume some robust segmentation algorithm (see Figure 2 for an example of colour-based segmentation using Statistical Region Merging, [16]) is available, that the walker operator or caregiver or other individual will be able to manually select segments belonging to the foreground of an initial frame I_0, and that the 3D model can be optimized so that its projection fits these segments.

Fig. 2. Example of colour segmentation with [16], original image (left) and segmented image (right).

Once the foreground is segmented and corresponding 3D model projection determined, four histograms (one per colour channel and one for texture) are constructed for each pair of leg segments. We consider pairs of segments as single regions to account for expected left-right symmetry.

To create colour histograms $HoC^{[k]}_{channel}$, $channel = \{red, green, blue\}$, each pixel for which $s(i,j,\overrightarrow{C}_{right,k}) = 1$ or $s(i,j,\overrightarrow{C}_{left,k}) = 1$ is classified

by colour channel value into one of 32 bins. Each $HoC^{[k]}_{channel}$ is normalized by the number of classified pixels $\sum s(i,j,\overrightarrow{C}_{right,k}) + s(i,j,\overrightarrow{C}_{left,k})$.

We create the texture histograms $HoG^{[k]}$ using similar methodology as for creating our colour histograms $HoC^{[k]}_{channel}$. To represent texture, we use Histograms of Oriented Gradients as described in [17]. I is converted to gray-scale and then two gradient filters are applied, resulting in a gradient magnitude and orientation measurement for each pixel in I. Each pixel for which $s(i,j,\overrightarrow{C}_{right,k}) = 1$ or $s(i,j,\overrightarrow{C}_{left,k}) = 1$ is is weighted by its magnitude and classified based on the its orientation into one of 64 bins. The histogram is normalized by the number of classified pixels. $\sum s(i,j,\overrightarrow{C}_{right,k}) + s(i,j,\overrightarrow{C}_{left,k})$.

At initial frame I_0 we compute templates $HoC^{[k]}_{true}$ and $HoG^{[k]}_{true}$, $k = 1,2,3$ using a manually positioned 3D model projection. We compute, for all subsequent frames, posterior likelihoods of state hypothesis $n = 1...N$ given observed colour and texture as follows:

$$\Delta^{[k][n]}_{HoC} = \max_{channel} \left| HoC^{[k]}_{true,channel} - HoC^{[k][n]}_{channel} \right| \quad (3)$$

$$P(I_{colour-model}|\overrightarrow{x}^{[n]}) = \lambda_{colour} exp\Big(-\lambda_{colour} \sum_k \Delta^{[k][n]}_{HoC}\Big) \quad (4)$$

$$\Delta^{[k][n]}_{HoG} = \left| HoG^{[k]}_{true} - HoG^{[k][n]} \right| \quad (5)$$

$$P(I_{texture-model}|\overrightarrow{x}^{[n]}) = \lambda_{texture} exp\Big(-\lambda_{texture} \sum_k \Delta^{[k][n]}_{HoG}\Big) \quad (6)$$

A. Appearance model evaluation

We compare the performance of specific colour and texture models to the four general cues in [13]. The general cues are summarized as follows (see [13] for implementation details). First and second, assuming homogeneity within these regions, then leg regions are characterized by a very low average gradient magnitude and uniformity in colour. Third is the expected symmetry that people tend to exhibit in their left and right body segments. While symmetry is assumed in our initialization of the appearance, it is explicitly evaluated in [13]. Fourth is the expected contrast between the foot and an uncluttered floor.

The likelihoods $P(I_{cue}|\overrightarrow{x}^{[n]})$ for a set of N state hypotheses ($n = 1...N$) is compared to two error measures: distance and orientation. For each cylinder k in the true image foreground $s(i,j,\overrightarrow{\mu}_x)$, a centroid coordinate $C^{[k]}_{true}$ and major orientation $O^{[k]}_{true}$ is calculated. These two attributes are also calculated for each set of cylinder projections of the state hypotheses. The distance and orientation errors are computed as follows:

$$error^{[n]}_C = \sum_{k=1}^{6} \left(C^{[k]}_{true} - C^{[k][n]} \right)^2$$

$$error^{[n]}_O = \sum_{k=1}^{6} \rho^{[k]}_{true} \left(O^{[k]}_{true} - O^{[k][n]} \right)^2$$

where $\rho^{[k]}_{true}$ is the projected length of cylinder k, $\rho^{[k]}$ 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.

3000 state hypotheses were randomly generated for two users shown in Figures 3(a) and 3(b). The search space was constrained laterally to within the rollator frame, and between 30cm and 130cm depthwise from the camera. True poses were manually segmented for each user, and are also shown.

(a) User A

(b) User B

Fig. 3. Model projection of the true poses for two rollator users.

For each set of estimates, distance and orientation errors were computed. Figures 4, 5, 6 and 7 show $P(I_{cue}|\overrightarrow{x}^{[n]})$ versus distance (left) and orientation (right) error for each cue for users *A* and *B* respectively. We would expect for a given cue that $P(I_{cue}|\overrightarrow{x}^{[n]})$ would decrease as $error_C$ and $error_O$ increase. Indeed this happens for each cue for users *A* and *B*. Clearly, the low-gradient-expectation cue, as well as the specific colour-model and texture-model cues have the smoothest, most consistent performance. The colour uniformity cue applied to user *A* in Figure 4(d) does not seem to discriminate at all between hypotheses with respect to distance and orientation error. Applied to user *B* in Figure 6(d), it only weakly promotes good hypotheses over poor ones. The other cues, although they show tendancies to promote good hypotheses, do not consistently reject poor ones. These results demonstrate the advantage of using specific appearance models tailored to the individual users, as opposed to general cues.

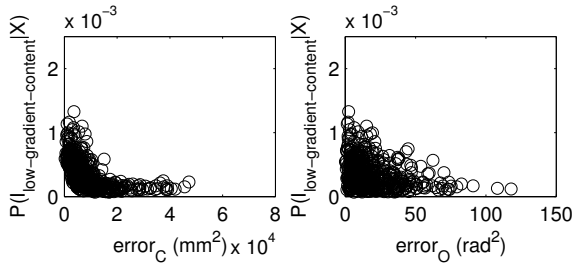
B. Combining weights

As mentioned in [13], the cues are not actully independent of one another given a hypothesis. We are exploring statistical dependencies between texture and colour using methods described in [18]. More specifically, we will in future augment the state with our specific texture and colour models, such that they are estimated rather than manually initialized. For now, we compute an overall likelihood of the state vector given the image cues as a product of the likelihoods given each cue:

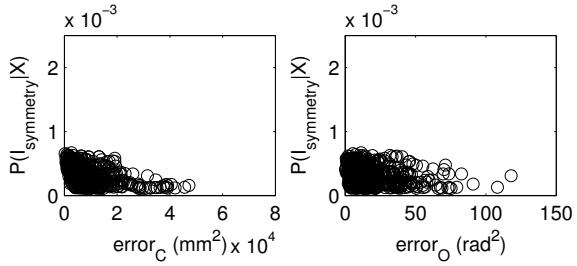
$$P(I|\overrightarrow{x}^{[n]}) = \eta_w \prod_{allcues} P(I_{cue}|\overrightarrow{x}^{[n]}) \quad (7)$$

where η_w is a normalizing constant.

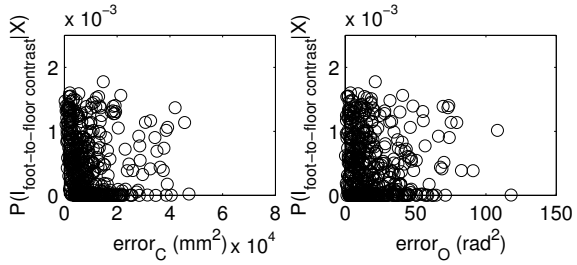
Figures 8 and 9 show $P(I|\overrightarrow{x}^{[n]})$ versus distance and orientation error for users *A* and *B* respectively. As we would expect, the distributions peak near $error_C = 0$ and



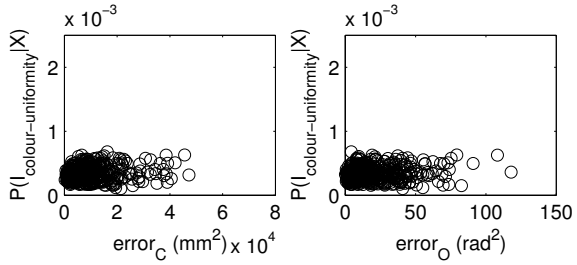
(a) Low-Gradient-Content



(b) Symmetry



(c) Foot-To-Floor



(d) Colour-Uniformity

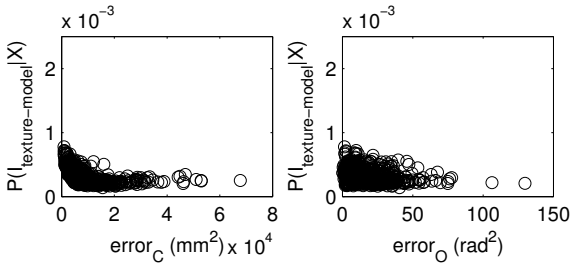
Fig. 4. Performance of cues from [13] for user A.

$error_O = 0$. However, the general cues result in more false-positives (hypotheses with larger errors associated with high posterior probabilities) than do the colour and texture cues conditioned on users *A* and *B*.

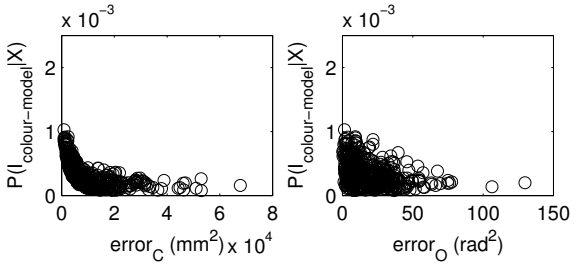
IV. TRACKING

As in [13], we chose a constant-velocity model for prediction because the motion of an elderly rollator user is typically slow and gradual.

The appearance model is initialized manually as described in Section III. We used the same "standing anywhere within walker frame" search criteria as in [13], but in future we plan to initialize the state based on the position of the segmented foreground. We apply the Condensation algorithm [19] to approximate the posterior probability given in Equation 2.



(a) Texture-Model



(b) Colour-Model

Fig. 5. Performance of proposed colour and texture cues for user A.

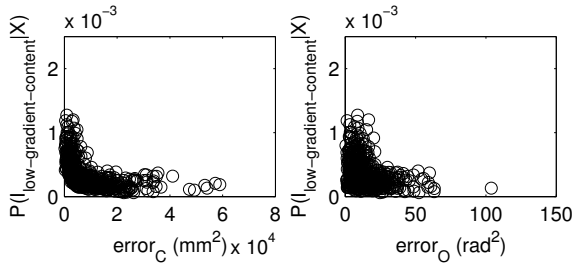
A. Preliminary comparison

We apply our appearance model (texture and colour) and tracking framework to a short 1.7-second video sequence featuring user *A*. Figure 10 shows these tracking results and results up to 1.7 seconds from [13] side-by-side for comparison.

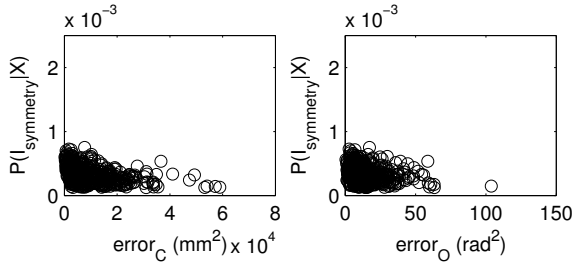
The results do indicate that our appearance model can coarsely infer relative depth. For each frame, the algorithm finds the correct depth of one lower limb relative to its counterpart and we can observe the general walking motion. Notice that the feet are better tracked in Figure 10(a) using the apperance model conditioned on the user. There remains much room for improvement. Knee and hip joint locations are still not estimated properly.

V. CONCLUSIONS AND FUTURE WORK

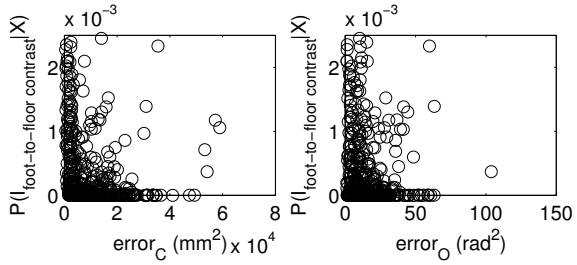
We have compared the use of specific colour and texture cues conditioned on the rollator user to more general cues describing legs of rollator users: homogeneity within leg regions indicated by a low gradient content and uniform colour, anthropometric symmetry, and contrast between the gray-level distributions of the floor and the feet. Each cue is evaluated separately against three images of different rollator users and for each user a set of 3000 hypotheses distributed within the operating space of the rollator frame. Although general cues tended to promote state hypotheses with smaller distance and orientation errors when combined, their individual performances (with the exception of gradient content) also promoted several hypotheses with larger distance and orientation errors. Colour uniformity by far was the weakest cue and, at best, rejected very poor state hypotheses. Specific cues are demonstrated as more effective and consistent in pose estimation than general cues. Further, they are less reliant on heuristic observations. In future we will therefore



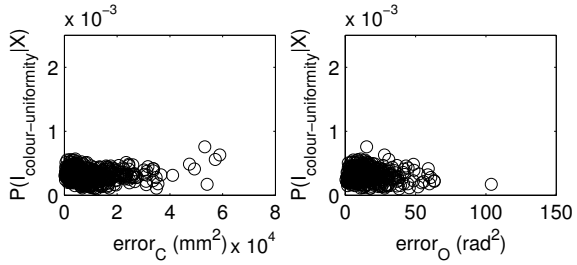
(a) Low-Gradient-Content



(b) Symmetry

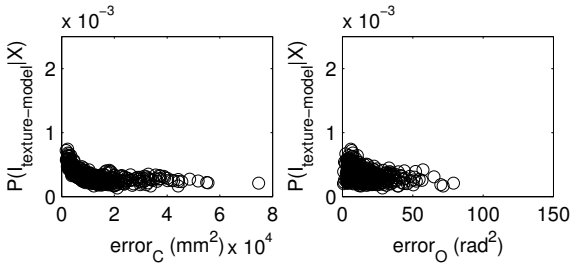


(c) Foot-To-Floor

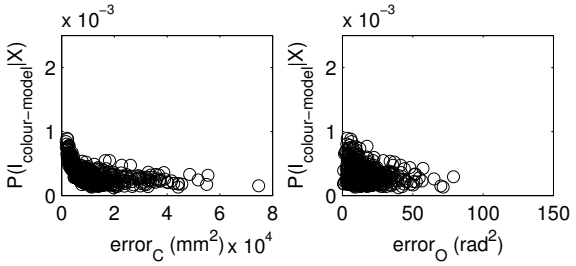


(d) Colour-Uniformity

Fig. 6. Performance of cues from [13] for user B.

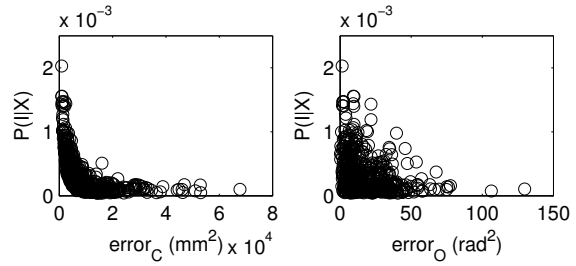


(a) Texture-Model

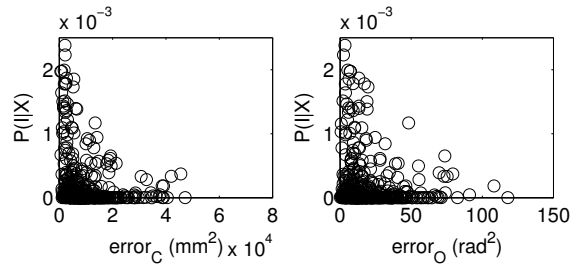


(b) Colour-Model

Fig. 7. Performance of proposed colour and texture cues for user B.



(a) Combined texture and colour cues conditioned on the user.



(b) Combined general cues from [13]

Fig. 8. Combined cue performance for user A.

aim to include texture and colour in our state estimation, rather than manually initializing them as templates as is done in this work. Preliminary tracking results for a short video sequence support the superior performance of cues conditioned on the rollator user. In addition to evaluating tracking over longer sequences, we intend to explore methods for handling occlusion and estimating joint position.

VI. ACKNOWLEDGMENTS

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.

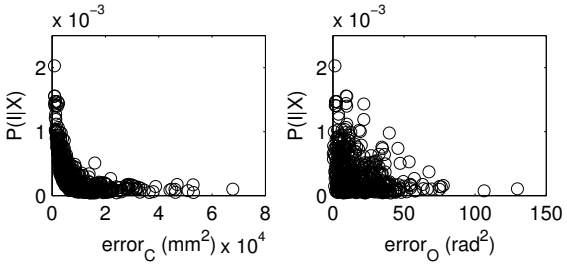
REFERENCES

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

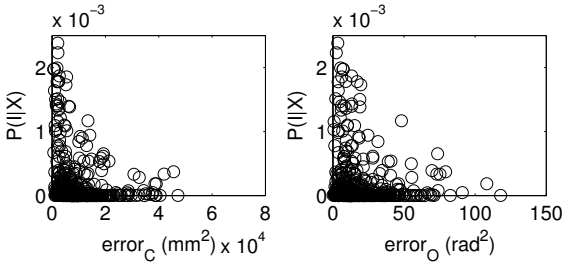
[2] J. Glover, D. Holstius, M. K. Montgomery, A. Powers, J. Wu, S. Kiesler, J. Matthews, and S. Thrun, "A robotically augmented walker for older adults," Carnegie Mellon University, School of Computer Science, Tech. Rep. CMU-CS-03-170, 2003.

[3] 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: IEEE Computer Society, 2008.

[4] Y. Hirata, A. Muraki, and K. Kosuge, "Motion control of intelligent walker based on renew of estimation parameters for user state," *Intel-*

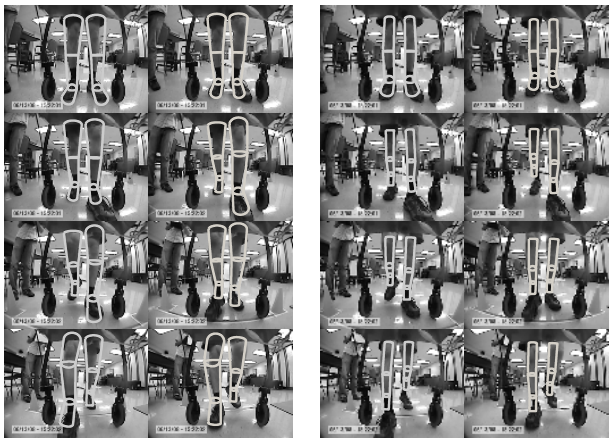


(a) Combined texture and colour cues conditioned on the user.



(b) Combined general cues from [13]

Fig. 9. Combined cue performance for user B.



(a) Manually initialized appearance model.

(b) General appearance model from [13].

Fig. 10. From left to right, top to bottom, tracking results for 10 sequential frames at 5fps. Cylinder projections indicate the *mean* of the posterior distribution.

ligent Robots and Systems, 2006 IEEE/RSJ International Conference on, pp. 1050–1055, Oct. 2006.

[5] M. Alwan, A. Ledoux, G. Wasson, P. Sheth, and C. Huang, "Basic walker-assisted gait characteristics derived from forces and moments exerted on the walkers handles: results on normal subjects," *Medical Engineering & Physics*, vol. 29, pp. 380–389, 2007.

[6] P. Sheridan, J. Solomont, N. Kowall, and J. Hausdorff, "Influences of executive function on locomotor function: divided attention increases gait variability in alzheimers disease," *Journal of the American Geriatrics Society*, vol. 51, no. 11, pp. 1633–1637, 2003.

[7] 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.

[8] J. Brach, J. Berlin, J. VanSwearingen, A. Newman, and S. Studenski, "Too much or too little step width variability is associated with a fall history in older persons who walk at or near normal gait speed," *Journal of NeuroEngineering and Rehabilitation*, vol. 2, no. 1, p. 21, 2005. [Online]. Available: http://www.jneuroengrehab.com/content/2/1/21

[9] H. Ning, "Kinematics-based tracking of human walking in monocular video sequences," *Image and Vision Computing*, vol. 22, no. 5, pp. 429–441, November 2004.

[10] 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*, vol. 104, p. 90126, 2006.

[11] D. Bullock and J. Zelek, "Towards real-time 3-d monocular visual tracking of human limbs in unconstrained environments," *Real-Time Imaging*, vol. 99, no. 7, pp. 323–353, November 2005.

[12] 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*, 1997, pp. 928– 934.

[13] S. Ng, A. Fakih, A. Fourney, P. Poupart, and J. Zelek, "Probabilistic 3d tracking: Rollator users leg pose from coronal images," in *Computer and Robot Vision*, 2009, accepted, to appear in.

[14] D. A. Winter, *Biomechanics and Motor Control of Human Movement*, 2nd ed. Toronto: Wiley, 1990.

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

[16] R. Nock and F. Nielsen, "Statistical region merging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1452–1458, 2004.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, p. 91110, 2004.

[18] C. Zhou and B. Mel, "Cue combination and color edge detection in natural scenes," *Journal of Vision*, vol. 8, no. 4, pp. 1–25, 2008.

[19] M. Isard and A. Blake, "Conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, p. 528, 1998.