Real-time articulated human body tracking using silhouette information

Anonymous

Abstract

In this paper we describe a system for visual tracking of the limbs of a human body. The system uses a 3D computer graphics model of the human figure and optimizes its parameters to fit the person's silhouette in one or more camera views. The system incorporates joint angle constraints and a novel method for dealing with ambiguous edges. The system operates in real-time on multiple views and has been evaluated on the CMU MoBo corpus with ground truth data and an automatically evaluated performance metric.

1. Introduction

The visual understanding of human actions is a burgeoning field with a wide range of applications, from motion capture for the film industry through surveillance to human computer interaction. In each of these areas there is a need to understand from images how a human body is moving. Many previous approaches to these problems have constrained the problem by either disregarding the articulations of the human body, treating it as a unitary, if protean, object [1]; or treating only a single aspect of the body, such as head pose, facial expression or hand gesture [2].

A number of applications however, require an approach with much more detail than the former, while modelling the whole body and parametrizing a wider range of motions than permitted by the more narrowly focused systems. Such applications include human-computer interaction, motion-capture for animation and biomedical applications, and surveillance. Clearly, some of these tasks might be better solved with special instrumentation or sensors, but the generality of a purely visual solution, together with its potential for low-cost deployment on standard, even existing, hardware; for application to legacy visual media (video indexing, sports video annotation); and the desire for a solution without intrusive instrumentation, for potentially covert use, and which parallels human abilities together make a computer vision approach desirable.

Surveillance, a rapidly growing area of computer vision research, will continue to demand more and more sophisticated modelling of people to progress from tracking of individuals to understanding of people's actions and interactions with objects and other people: Is this person shoplifting? Is that person drawing a weapon? Are those people shaking hands? Recently the Human Identification at a Distance Project has also spurred activity in identifying individuals by gait, though many approaches extract generic silhouette features without attempting to track the articulated motion of the body [16].

In this paper we propose a novel method of tracking articulated human bodies in video data and evaluate our algorithms ground truth for a standard dataset. The following section reviews previous work in the field. In Section 3 we describe our model and explain in Section 4 how it is used to track video data. Section 5 describes our approach to performance evaluation and Section 6 describes the experiments that have been conducted to test the algorithms described.

2. Previous work

Several researchers have started to tackle the problem of articulated body tracking since early work by Hogg [14]. Gavrila [11] reviews work in this field, though there has been a growth of interest in recent years. Because of the complexity of the problem and the large quantities of data involved, previous work has been limited by computer power. This has also meant that little evaluation has taken place, with performance measures being limited to subjective evaluation and the publication of a few sample frames.

2.1. Tracking with direct methods

Bregler and Malik [6, 7] describe the tracking of an articulated body by fitting a three-dimensional model to multiple views, using the direct methods derived from the tracking described by Bergen *et al.* [4]. The model is used to predict the motion of each part of the image in terms of the changes in the parameters of the model, which is represented with exponential twists which we describe in Section 4.2. Given a frame I of data at time t, when the model is assumed to be correctly fitting the person with parameters, θ, a new image received at time $t + 1$, the old frame can be rerendered as it would appear for a hypothesized change $\dot{\theta}$ in the parameters. Their system tries to minimize the least-squares intensity error between the new frame and the warped old frame by solving the nonlinear least squares problem set up by representing the derivative of the brightness constancy equation:

$$\nabla I d\mathbf{x} + \frac{\partial I}{\partial t} dt = 0 \tag{1}$$

recast in terms of changes in the parameters.

Cham and Rehg [8] track a two-dimensional figure in monocular data using a probabilistic model to estimate the likelihood of a given parameter change in a multi-hypothesis framework.

2.2. Fitting with edges

Drummond and Cipolla [10] have used a similar model and optimization framework to Bregler and Malik with a different tracking method. They too have a computer graphics model, conceived as a hierarchy of rigid parts with joints represented as exponential twists, in this case with eighteen degrees of freedom. The solids used for the body parts are modelled as intersections of pairs of quadrics.

The main difference with the work of Bregler and Malik, is that the features used for fitting the model are edges predicted in the model, which are forced to conform to image points with high intensity gradient. From the computer graphics model, the locations of body part edges are predicted. These predicted edges are sampled at regular intervals, and perpendicular to each one, a search is carried out to find the local maximum in the image gradient, which is hypothesized as the true location of the edge.

Since there are many fewer edge features than image pixels on the body, and since there is no image warping step, this system is able to run in real time (25Hz) on a single view.

Gavrila and Davis [12] used a computer graphics model based on a 22-parameter stick figure which was "fleshed out" with hand-crafted superquadric limbs. These models were tracked automatically in four views (back, front, left and right) by matching the edges in the image (with background subtracted) to the predicted edges of the model using chamfer matching. Because of the high dimensionality of the search space, it was decomposed with fitting proceeding by first aligning the torso, then the limbs. This system is also able to initialize the model parameters when a sequence begins with a pose with distinctive silhouette.

More recently, Poon and Fleet [19] and Lee *et al.* [15] have described methods for fitting an articulated body model to video of human motion, using particle filtering approaches, which sample particles from posterior probability distributions across parameter space. The former track edge features in monocular sequences, with hand initialization of the models, but quote processing times of up to 7 minutes per frame. The latter use three views, fitting their model to the segmented foreground region in each, and have designed an automated system for initializing the model parameters. Plaenkers and Fua [18] use a combination of edge and region information to track people in stereo data with a model based on metaballs.

3. Model

The system for human body tracking that we have developed has a computer graphics model which consists of a hierarchy of rigid parts whose relative positions are parametrized by exponential twists [7]. The model is of variable resolution allowing detailed modelling when more information is available (when part of the person is close to the camera, in the field of view of a telephoto camera, or is visible to multiple cameras) and of lower resolution when less visual information about a body part is available, or when an application requires less detailed information.

The three-dimensional model used consists of ellipsoids and generalized cylinders. Ellipsoids are parametrized by the lengths of their three principle axes. The generalized cylinders are the frusta of elliptical cones, parametrized by a height and the lengths of major and minor axes of the elliptical cross-section at one end, with a relative magnification applied equally in both dimensions at the far end.

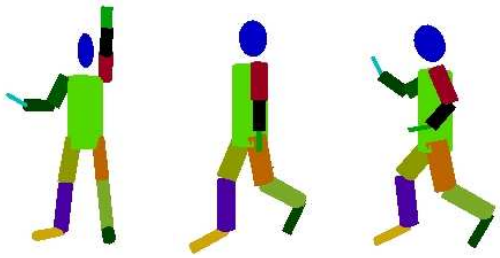

Figure 1: Three views of the model showing a range of poses.

The body model has up to fourteen rigid components: torso, head, upper & lower arms, hands, thighs, calves and feet. The head is the only component for which an ellipsoid is used. Figure 1 shows the fourteen-component model in various poses. In the experiments described here, the model has a restricted set of degrees of freedom, sufficient to capture the most significant movements in a walking person. The dimensions are not allowed to vary, and only a restricted subset of the angular twist parameters are tracked.

4. Tracking

The system tracks articulated motions using the sequence of processes shown in figure 2 on each new frame of data. An estimate of the model parameters is derived from the parameters for the previous frame. (In practice a constant angular velocity model has not been found to be useful, so the old parameters are used directly as an initial estimate.) Given this model pose, in each view, the occluding contours of the limbs are calculated, and correspondences to these contours are sought in the silhouette for that view, calculated by background subtraction [3]. A new set of parameters

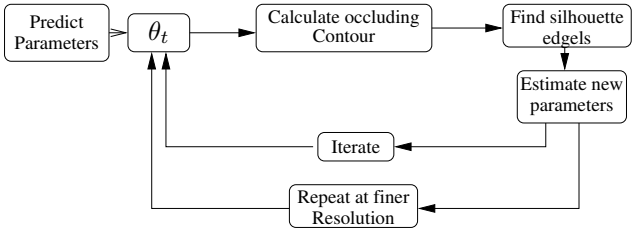

Figure 2: A schematic view of the operations involved in parameter estimation.

is estimated, and the process repeated. We apply the procedure in a course-to-fine search (typically with two scales separated by a factor of two) to achieve faster convergence over a larger search space, with several (typically three) iterations at each scale.

4.1. Silhouette features

Drummond and Cipolla [10] fit the predicted edges in the graphical model to edges found in the video image, masking out edges not on the person to be tracked with a background subtraction procedure. Here we choose to fit the model to the more well-defined silhouette of the model, also calculated with background-subtraction.

The model with the current pose parameters is projected onto the image plane to find the occluding contour of the model, as shown in Figure 3. The projections of the curved sides of each body part are always tracked, and individual body parts are flagged if either end should be tracked. The projected edges are sampled at regular intervals (typically every pixel) and within some window (typically 8 pixels at any given scale) along a normal at each sample point, a silhouette edge (a foreground-to-background transition) is searched for. This is shown in Figure 4.

While both are subject to failures of background subtraction, silhouette edges have two advantages over image gradient features in that texture on the tracked object can generate false edges, and the sign of the transition is known (foreground to background), whereas the intensity edge could be light-to-dark or dark-to-light. Furthermore the normal only needs to be searched in one direction since if the sample point is in the foreground, the search must be in the direction away from the object, but if the sample point is already on the background, the search should be towards the inside of the limb.

4.2. Twist representation of kinematic chain

The model is parametrized using the exponential twist formulation from the robotics literature [17] that is used by Bregler and Malik [7] and Drummond and Cipolla [10] to describe the pose of each part with respect to its 'parent' parts. We briefly review the derivation of image motions

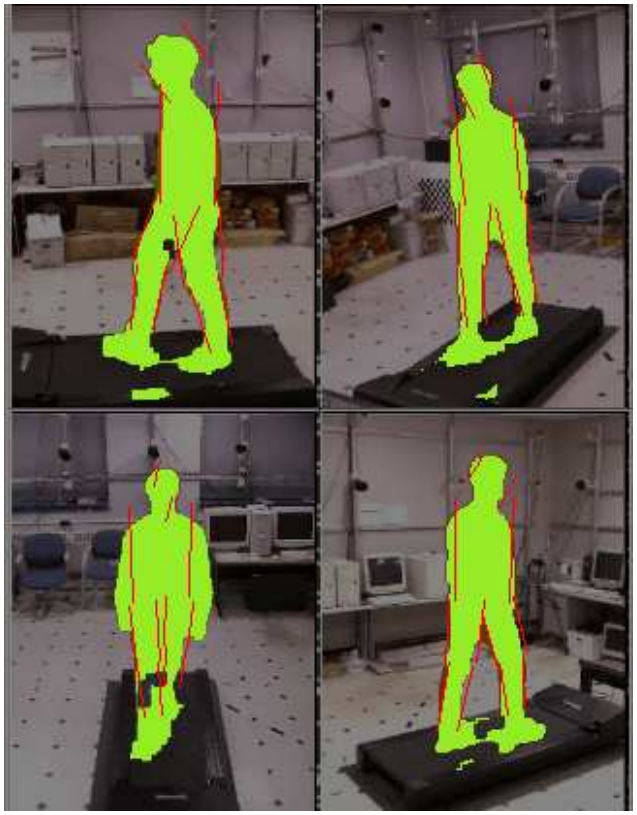

Figure 3: The foreground regions with the predicted model edges superimposed for each view.

from parameter changes in the twist representation. Each joint k in the model is a revolute joint about an axis ω, with a point $\mathbf{q}$ on the axis. For a rotation θ_k, this is represented with a twist

$$\xi_k = (v_1\, v_2\, v_3\, \omega_x\, \omega_y\, \omega_z)^T \quad (2)$$

$$= \begin{pmatrix} -\theta_k \omega \times \mathbf{q} \\ \theta_k \omega \end{pmatrix} \quad (3)$$

This rotation has homogeneous transformation $e^{\hat{\xi}_k}$ where:

$$\hat{\xi}_k = \begin{pmatrix} 0 & -\omega_z & \omega_y & v_1 \\ \omega_z & 0 & -\omega_x & v_2 \\ -\omega_y & \omega_x & 0 & v_3 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (4)$$

Now, a kinematic chain of rigid limbs $1, \ldots, K$ joined by revolute joints can be represented by a product of such transformations. A point $\mathbf{q}_o$ on joint K is

$$\mathbf{q}_c = \left(\prod_{k=1}^{K} e^{\hat{\xi}_k} \right) \mathbf{q}_o \quad (5)$$

in the global coordinate system in which the chain is anchored. It can be shown [17] that if the joints are rotating

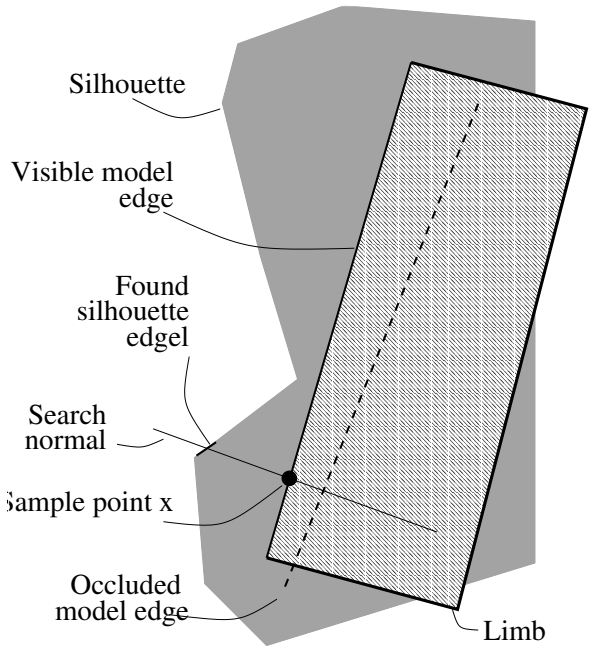

Figure 4: A model edge (full line) with one sample point and its search normal and a found silhouette edgel. The occluded model edge (dashed line) could also match against this silhouette edgel.

with angular velocities $\dot{\theta}_k$ then the velocity in the global coordinate system is:

$$\dot{\mathbf{x}} = \left(\prod_{k=1}^{K} \hat{\xi}'_k \right) \mathbf{q}_c, \tag{6}$$

where

$$\hat{\xi}'_k = \mathbf{Ad} \left[\left(\prod_{k=1}^{K-1} \hat{\xi}_k \right) . \xi_K \right], \tag{7}$$

and where

$$\mathbf{Ad} \left[\begin{pmatrix} R & p \\ \mathbf{0} & 1 \end{pmatrix} \right] = \begin{pmatrix} R & \hat{p}.R \\ \mathbf{0} & R \end{pmatrix}, \tag{8}$$

denoting the adjoint (6×6) matrix of a homogeneous transformation.

4.3. Fitting

Each silhouette edgel found by searching along a normal gives a hypothesized displacement (dx, dy) for the point $\mathbf{q}_c$ on the model's occluding contour. In this formalism, tracking consists of finding the set of parameter changes $\dot{\theta}$ that minimizes the sum-squared displacement. Since we know the model point's position, Equation 6 gives us the displacement induced by a set of parameter changes. Separating this into x and y components:

$$\begin{aligned} H_x(\mathbf{x}).\dot{\theta} &= dx \\ H_y(\mathbf{x}).\dot{\theta} &= dy. \end{aligned} \tag{9}$$

Combining the displacement equations 9 for all pixels, we arrive at:

$$H.\dot{\theta} + \mathbf{d} = \mathbf{0}. \tag{10}$$

When we have multiple views, each pixel in each view contributes two rows to a single equation of this form. Equation 10 is solved by least squares:

$$\dot{\theta} = -(H^T.H)^{-1} H^T.\mathbf{d} \tag{11}$$

Since the equations are in fact nonlinear, the process is iterated.

4.4. Ambiguous edges

It should be noted that there are often self-occlusions by the limbs of a human body, and this leads to ambiguity in some of the edges, as shown in Figure 4. Here the found silhouette edge could belong to the unoccluded edge, or it could belong to the occluded edge as it becomes disoccluded. All found silhouette edgels are compared against the predicted model edges of the other limbs, and if such ambiguity is found, then the contribution of the edgel is diminished and shared between the two hypothesized edges. If the found edgel lies a distance d_a and d_b from model edges a and b respectively, then the edgel contributes to a's equation with weight $\frac{d_b}{2(d_a+d_b)}$.

4.5. Constraints

When tracking an object with a model with many degrees of freedom, it is important to introduce as many constraints as possible to limit the exploration of false local minima in an optimization formulation, and reduce the search space in a problem formulated as a search. Drummond and Cipolla [9] discuss the application of constraints of the type of allowable motion in an articulated body (*e.g.* hinge or slide). In our implementation such constraints are enforced simply by determining which of the twist parameters are included in the optimization and which are not. However a further constraint can be applied by limiting the range of joint angles to match those physically possible in a normal human, such as a knee bending backwards but not forwards. We introduce such joint angle constraints by specifying maximum rotations for the joints (independently for each of the three Euler angles) and adding further terms to the optimization problem to prevent solutions where these are violated.

A penalty proportional to the square of the angle by which a joint exceeds a limit is applied.

Naturally much more complex constraints can be created with non-separable constraints on the 3D rotation of a joint; the requirement that the body parts do not intersect one another or objects in the environment (particularly the ground plane). Further, for many types of motion it is known that the feet must be in contact with the ground plane, with no motion relative to it, and the problem could be further constrained by insisting on "typical" poses or motions [20, 21]

4.6. Initialization

In the experiments conducted so far, the body model is approximately adjusted by hand to match the limb lengths and pose in the first frame of video. Alignment with the centroid of the foreground regions can be used to initialize the overall model translation, but more sophisticated algorithms [15] must be used to initialize the pose and limb lengths, or multiple different poses can be hypothesized and tested.

5. Evaluation

To evaluate the performance of our tracking algorithm, we have tested it on a publicly available dataset—the CMU MoBo corpus [13] that was created as part of the Human Identification at a distance project. This database provides 30fps video sequences with six simultaneous calibrated views of each of 25 subjects walking on a treadmill in each of four conditions (slow walk, fast walk, carrying a ball, incline). A foreground segmentation is provided with each frame of data, derived from an automatic background subtraction algorithm. The database is proposed for the development of gait recognition algorithms [16]. For the evaluation used here, we have used six subjects (04002,04006,04011,04022,04037,04068) in the `slowWalk` and `fastWalk` conditions, with up to four of the views (vr03, vr05, vr07, vr13: camera positions are shown in Figure 5). We have down-sampled the 476×640 to 288×320.

To evaluate the performance of a tracking algorithm, we need to define one or more performance criteria. Previous systems have made do with subjective evaluation by the authors of whether tracking works or not for a sequence, and the quantity of data treated in previous work has been very small. When motion capture data is available [5] a Euclidean distance measure for the location each of the tracked joints can be used, and could be summed into a global measure for the whole body. Additional error measures can be evaluated using other parameters such as joint angles. However, the presence of motion capture equipment on the subject may also interfere (or help) with the visual tracking.

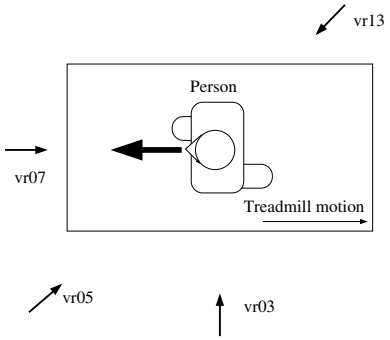

Figure 5: The orientations of the cameras in a plan view of the CMU MoBo treadmill setup.

5.1. Joint position error

For the CMU MoBo corpus, motion capture data is not available, so we have marked up several of the sequences with the positions of certain joints. In practice the most interesting information for this data is the position of the feet. Further, fiducial points on the feet are easier to define, so foot positions have been used for this evaluation. The points that we have labelled (in the original 468×640 images) are the estimated position of a point in the centre of each ankle joint. This is necessarily approximate, and the errors that we quote include some component due to errors in markup, but the relative performance of the algorithm with different parameters should be significant. In the tables below, we show the mean squared error in the foot position averaged across both feet and all marked frames, and averaged across al the views used for tracking. Each sequence consists of 340 frames marked up at least every 10 frames.

5.2. Silhouette fit error

In addition to the joint position error above, we propose a simpler error measure for preliminary evaluation of the tracking system, which does not require manual ground truth data to be generated. This error measure, E_S, can be summed up as the classification error when using the model to determine which pixels are part of the object and which are not:

$$E_S = \frac{||(S_M \backslash S_T) \cup (S_T \backslash S_M)||}{||S_T||}, \qquad (12)$$

where S_M is the set of pixels classified by the model as being foreground, and S_T is the set of pixels labelled as foreground in the ground truth. In practice no definitive ground truth labelling is available, but we use the background subtraction masks provided with the CMU MoBo data as though they were correct labellings. This error measure is necessarily view-based and we evaluate it for each

camera available. Results are averages across all frames in all four views, regardless of the number of views being used for tracking.

It can be seen that E_S will be low when the silhouette of the model closely matches the silhouette in the background subtraction labelling. Since we use this same silhouette in our fitting procedure, this could be considered a very artificial method, and it would clearly be easy to "cheat" with a more general model. However, given the very specific nature of the model, we believe that E_S provides a good evaluation criterion for initial studies. It does not penalize poor tracking within the silhouette, but such poor tracking is penalized in other views or in subsequent frames when the silhouette changes.

6. Experiments

For evaluation of the tracking algorithms, we began with a simple model with only six components and seven degrees of freedom. The components are torso, head, thighs and calves, with only the global translation (x, y, z) and the forward-backward rotation of knee and hip joints were allowed to vary. These degrees of freedom capture the principal variations of interest when recognizing gait.

Table 1 shows tracking performance for the algorithms using two, three or four views (cameras 03 and 13, then adding 05 and 07) for the `slowWalk` and `fastWalk` sequences.

Views	Iterations	Data set	E_s (%)	Joint position (pixels)
2	3*	`slowWalk`	21.2	18.3
2	3	`slowWalk`	20.6	12.9
3	3	`slowWalk`	17.6	8.7
4	2	`slowWalk`	17.5	8.8
4	3	`slowWalk`	17.5	8.8
4	5	`slowWalk`	17.5	8.9
4	6†	`slowWalk`	17.8	9.6
2	3*	`fastWalk`	22.7	25.4
2	3	`fastWalk`	21.0	10.4
3	3	`fastWalk`	18.9	8.9
4	2	`fastWalk`	18.9	11.0
4	3	`fastWalk`	18.7	8.8
4	5	`fastWalk`	18.7	8.5

Table 1: Performance evaluation of the tracking algorithms. Conditions marked * are without ambiguity de-weighting of Section 4.4, that with † uses only a single scale.

As the number of views increases, it can be seen that the accuracy of the fit improves. Similarly it can be seen that the process of de-weighting ambiguous edgels improves the accuracy, (with a 10% speed penalty involved in searching for the ambiguity). Three iterations per scale seem to suffice, but a multiscale approach is better than fitting at a single resolution. The two performance measures used seem to correlate well, though there is more discrimination in the scores derived from the hand-labelled data than from the silhouette match score S_E.

Table 2 shows the average processing time per frame and indicates that with two or three views the tracking can be carried out in real time at the original frame rate of 30 frames per second. While here we are not computing the background masks which would add some overhead, there are several improvements to the code that could be made to reduce the times presented here without affecting the accuracy.

Views	Iterations	Time (ms)
2	3	31
3	3	38
4	2	33
4	3	44
4	5	62

Table 2: Performance evaluation of the tracking algorithms. Times are average processing time per frame for a dual 2.8GHz Pentium 4 machine. "Iterations" is the number of iterations at each of two scales.

Figure 6 shows the degree of alignment between the model tracking and the ground truth, showing the x coordinate for both feet over one sequence that was ground truthed every third frame. Qualitatively, the tracks follow the ground truth closely, though the peaks are not as high as in the ground truth, probably because the model limb lengths are incorrect. This substantiates our qualitative evaluation that tracking is "correct" for all these sequences—while limitations in the model mean that the fit is not perfect, the limbs are correctly tracked without the tracker losing the subject or confusing left and right legs. (See supporting video submitted with paper.)

7. Conclusions

In this paper we have demonstrated a real-time multi-view articulated human body tracking system that tracks with high accuracy. Tracking performance has been evaluated on a publicly available dataset, and with objective performance criteria.

In the future we plan to extend our evaluation to more of the CMU MoBo dataset and to add automatic initialization, together with adaptation of limb dimensions to the observed data. We will evaluate the tracking of more degrees of freedom and with more complex motions.

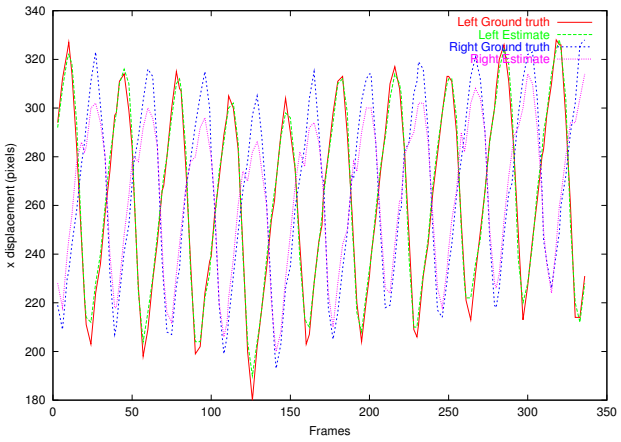

Figure 6: A graph of the x-coordinate of the left and right feet as marked by hand and as tracked automatically. (subject 04002, side view vr03, `slowWalk`)

References

[1] *IEEE Workshop on Performance and Evaluation of Tracking and Surveillance Systems*, 2002.

[2] *International Workshop on Face and Gesture Recognition*, 2002.

[3] Anonymous. Reference removed to preserve anonymity. In *A Conference*, recently.

[4] J.R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In G. Sandini, editor, *European Conference on Computer Vision*, number 588 in LNCS, pages 237–252. Springer-Verlag, 1992.

[5] A.F. Bobick and A.Y. Johnson. Gait recognition using static activity-specific parameters. In *Conference on Computer Vision and Pattern Recognition*, 2001.

[6] C. Bregler. Learning and recognizing human dynamics in video sequences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico*, pages 568–574, 1997.

[7] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Conference on Computer Vision and Pattern Recognition*, 1998.

[8] T.-J. Cham and J.M. Rehg. A multiple hypothesis approach to figure tracking. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, June 23-25*, pages 239–245, 1999.

[9] T. Drummond and R. Cipolla. Real-time tracking of multiple articulated structures in multiple views. In *European Conference on Computer Vision*. IEEE, 2000.

[10] T. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *International Conf. on Computer Vision*. IEEE, 2001.

[11] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision, Graphics and Image Understanding*, 73(1):82–97, January 1999.

[12] D. M. Gavrila and L. S. Davis. 3-D model-based tracking of humans in action: A multi-view approach. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco*, pages 73–80, 1996.

[13] R. Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Carnegie Mellon University, June 2001. `http://hid.ri.cmu.edu/HidEval/index.html`.

[14] D. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[15] Mun Wai Lee, Isaac Cohen, and Soon Ki Jung. Particle filter with analytical inference for human body tracking. In *IEEE Workshop on Motion and Video Computing*, pages 159–165, December 2002.

[16] Yanxi Liu, Robert Collins, and Yanghai Tsin. Gait sequence analysis using frieze patterns. In *European Conference on Computer Vision*, number 2351 in LNCS, pages 657–671. Springer-Verlag, 2002.

[17] R. M. Murray, Z. Li, and S. S. Shastry. *A mathematical introduction to robotic manipulators*. CRC Press, 1993.

[18] R. Plaenkers and P. Fua. Model-based silhouette extraction for accurate people tracking. In *European Conference on Computer Vision*, number 2351 in LNCS, pages 325–339. Springer-Verlag, May 2002.

[19] Eunice Poon and David J. Fleet. Hybrid Monte Carlo filtering: Edge-based people tracking. In *IEEE Workshop on Motion and Video Computing*, pages 151–158, December 2002.

[20] H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *European Conference on*

Computer Vision, volume 1 of *LNCS*, pages 784–800. Springer-Verlag, May 2002.

[21] Ying Wu, J. Lin, and T.S. Huang. Capturing hand natural articulation. In *International Conf. on Computer Vision*, volume II, pages 426–432, 2001.