M.I.T. Media Lab Vision and Modeling Group Technical Report No. 223

Analyzing and Recognizing Walking Figures in XYT

Sourabh A. Niyogi and Edward H. Adelson

sourabh@media.mit.edu and
adelson@media.mit.edu
Perceptual Computing Section
MIT Media Laboratory
20 Ames St., Cambridge, MA 02139

1 Introduction

We have derived a novel approach to the detection and recognition of human gait. In *gait detection*, we find a spatiotemporal pattern that signals the presence of a walking person. In *gait recognition*, we seek to identify the individual who is walking. It is known that humans can detect and recognize gait with reduce spatiotemporal sequences (such as moving light displays) [3, 10], and we would like to give similar capabilities to machines.

Any reasonable approach to the interpretation of human motion must impose a model of a human and explain how visual observations are to be fitted to the model. Model recovery is difficult for a number of reasons. As with most object model recovery, this process should be insensitive to lighting, position, and size. In modeling humans, the recovery process should not be sensitive to clothing or any other features specific to a particular individual. Furthermore, unlike most objects, the human body is composed of a large number of parts which can move non-rigidly with respect to one another. Since only some of the parts are visible at any given time, any approach that attempts to recover a full three-dimensional model will have considerable difficulty defining the position of occluded body parts. Some make the problem more tractable by interpreting motion with marked feature points [19, 12, 13, 1, 4]. Recovering these features to these models is not a trivial task, and there have been several attempts to recover models from real imagery, each with a different goal [7, 8, 16, 18, 17, 14, 5, 6, 15, 20].

2 Overview

Our approach to human motion analysis takes a novel approach to model recovery, based on the observation that walkers generate special signatures in space-time. We analyze the patterns and use them to estimate the parameters of a simple stick-figure model.

Figure 1 shows a image sequence cube of a frontoparallel walker. This "cube" is formed by stacking each of the frames in an image sequence one right after another. An XT-slice of the cube near the walker's ankle reveals a unique braided signature for walking patterns. Figure 2 shows an XT-slice obtained near the walker's head; the XT-slice indicates that the head undergoes pure translation during normal walking. Figure 3 shows the walker's two legs criss-crossing over one another as the walker walks from left to the right. These braided patterns are generated by all human walkers. Figure 4 shows three additional image sequence cubes for different people and different locations.

The approach we take to *detect* gait is to find translating blobs in image sequences, and test if the XT-slice of the lower half of the blob contains a gait signature. If a signature exists, one can be reasonably certain that a human walker generated such a pattern. The algorithm can then proceed to model the walking pattern of the individual.

The approach we take to *model* gait is to recover a set of contours for these XT-slice signatures. We recover not just the spatiotemporal edges for the XT-slice taken at the ankle, but for all XT-slices in the translating blob, effectively tracking the contours of the walker as he walks in space-time. Given these contours, it is fairly straightforward to produce a stick model of a person. And, from a stick model of an indi-

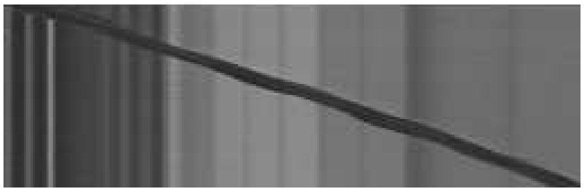

Figure 2: XT-slice of image sequence near head.

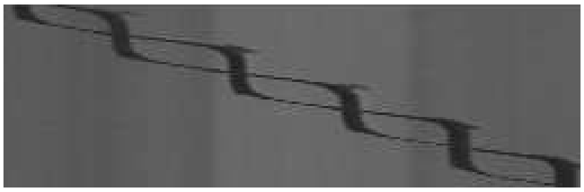

Figure 3: XT-slice of image sequence near ankle.

vidual's walk, we can try to recognize the individual.

We have been able to recognize gait from image sequences with a set of reasonable assumptions: (1) camera is fixed (2) the walker is walking at roughly constant speed (3) the walker is walking roughly frontoparallel relative to the camera (4) the walker is not camouflaged, or carrying anything that would obscure his braided gait pattern.

The overall scheme is shown in Figure 5. Roughly, the algorithm is as follows:

1. *Gait detection.* An image sequence may or may not contain a human walking. The frontoparallel walker can be coarsely modeled as a translating rod. If we can find rod-like translating objects, then we can test if the lower half of the object contains a braided walking pattern. Because the braided walking pattern is structured, we can use template matching for coarse recognition. If the template match is not a reasonable fit, the translating blob is not a walker. Otherwise, template matching results in two template signals, one signal representing the x-coordinate of the center of the left leg and another signal for the center of the right leg. See Figure 9 for an template overlaid on a braided pattern.

2. *Body tracking.* Once gait has been detected, we recover the frontoparallel walker's body contours with "snakes"[9]. Snakes are used to recover the four spatiotemporal edges of the braided pattern using the template match as a coarse guess. Figure 10 shows four recovered spatiotemporal edges overlaid on the XT-slice. The spatiotemporal edges recovered at the ankle are used to recover spatiotemporal edges at other heights. The end result is that the snakes accurately recover the moving contour of a person over time. See Figure 12 for the snake output of an image sequence.

3. *Gait modeling.* The recovered body contours are used to build a stick model of the walker. These four contours are averaged into two skeletons, and snakes are used to obtain refined estimates of the spatial positions of the hip, knee, and ankle. Line-fitting operations between these positions yield four angle signals for an image sequence. Figure 13 shows an example of a body contour fit; Figure 14 shows an example of one initial skeleton fit; Figure 15 shows a final skeleton fit; Figure 16 shows a stick model fit, for just one frame in an image sequence.

4. *Gait recognition.* The angle signals that define the stick model of the walker, such as those shown in Figure 17, are classified with a table of previously observed gait signatures. A standard k-nearest neighbor approach is sufficient to obtain recognition rates well above chance performance.

We construct a five-stick model of a walker through spatiotemporal analysis. This is not all of the information there is to gait. The free motion of the arms, the head, and feet are not modeled. This information cannot be recovered through direct spatiotemporal analysis, unlike analysis of the the braided pattern near the legs.

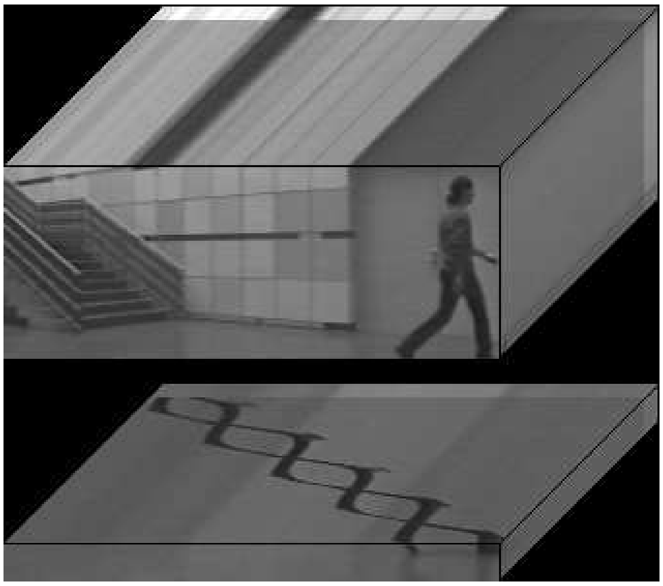

Figure 1: XYT Image sequence volume of a frontoparallel walker, sliced at the ankle. This is obtained by stacking frames of an image sequence to form a cube, where t is out of the page. Notice braided pattern in the XT-slice. A shoe appears occasionally at this particular height.

However, with the information that can be recovered, we show that there is sufficient information to obtain promising recognition rates.

We also obtain non-gait information, which if processed further, could be used for other sources of recognition information. The stick length ratios to the feature vector and the body contours themselves give considerable form information. One might be use such information to recognize men from women, or distinguish overweight people from the normal.

We now describe the algorithm in more detail.

3 Gait detection

Gait detection is solved by finding translating objects in an image sequence and testing if they contain a braided pattern in the lower half of the translating object. Moving objects are highlighted using a change detection operation between each image and the background. All moving objects will be highlighted in the XYT image cube; translating frontoparallel walkers form a plane in the XYT image cube, while XT-slices near the head reveal a line. One reason for this approach is that standard optical flow algorithms fail in regions where there are multiple motions, occlusion, and non-rigidly moving areas; many other human body tracking efforts use a change detection operation as well to bypass these problems[18, 7, 17, 16]. The constant background assumption employed here need not be so strong. A background appears as vertical stripes in an XT-slice. If the background is in motion, then the background will appear as oriented stripes in the XT-slice. Any scheme which is able to estimate background motion will be able to shear the XT-slice as if the background was constant.

The specific change detection algorithm we use transforms an n-frame image sequence $I(x,y,t)$ into another image sequence $O(x,y,t)$ by first computing a background $B(x,y)$ using median filtering and with that a variance $V(x,y)$. These are used to compute a new "change detected" image sequence which effectively highlights significant deviations from the background. One

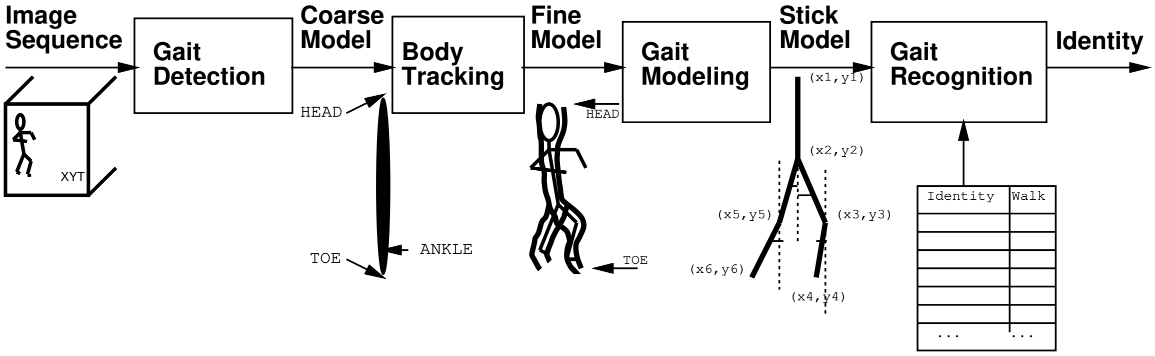

Figure 5: Gait information processing architecture.

can take a slice of the $O(x, y, t)$ cube to obtain change-detected XT-slices. The change detection algorithm run on the two images shown in Figure 2 and Figure 3 results in images shown in Figure 6 and Figure 7. Using robust statistics, one can recover the parameters which define the plane in XYT, or equivalently, the line of the XT-slices near the head. However, since we only had one walker in our image sequences, a simple regression technique was sufficient. The slope and intercept of the line in the XT-slice correspond to the walker's speed v and initial position x_0.

In order to detect that the translating object is a walker, we need to decide whether the XT-slice signature in Figure 7 is a braided pattern. Because the data is quite structured, we can correlate a small number of templates to the potential braided pattern. The template model is composed of three variable parameters, an amplitude A, a period T, and a skew p, in addition to the fixed parameters v and x_0. The template model is diagrammed in Figure 8, and is essentially two signals, $l(t)$ and $r(t)$, which can be correlated with each change-detected image $O(x, t)$ with a standard correlation measure. The best template match is found by searching for the maximum correlation over a small number of amplitudes, periods, and skew parameters. The range of amplitudes (A) that should be tested depends on the stride length and the distance of the walker to the camera. The range of periods (T) that should be tested depends on the walking speed of the person, the stride length, and the length of walker's body parts. If the best correlation is low, there is not a braided pattern, so we can reject the hypothesis that the translating blob is a walking person. If the correlation is high, then there is probably a braided pattern, so we may continue processing. The template match for the ankle XT-slice is shown in Figure 9. These two signals $l(t)$ and $r(t)$ yield a rough estimate of the centers of the both ankles as a function of time. Templating matching is done only once in processing an image sequence.

4 Body tracking

Initial snaking on XT-slices. Once we have detected a human gait pattern, we refine our rough estimate of the walker's pattern with "snakes"[9]. Snakes are splines which possess an internal energy, defined by their configuration, and an external energy, defined by an image energy function. Given an initial list of points that define a snake, the list of points will "climb" to the local maxima in the energy function. The two signals obtained from the template match are used to initialize two snakes, $l(t)$ and $r(t)$; The energy function used is just the change-detected XT-slice, i.e. the one shown in Figure 7. If the template match yields a correct answer, using that energy function attracts both snakes to the center of each ankle. Letting the snakes settle on

Figure 4: Three additional image sequences of frontoparallel walkers. Again, notice braided pattern in the XT-slice near the ankle in each image sequence.

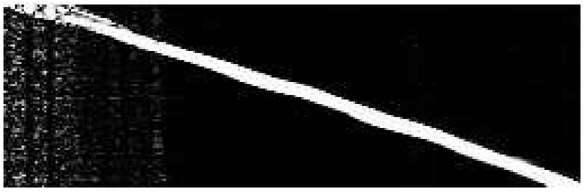

Figure 6: XT-slice near head run through a change detection operation.

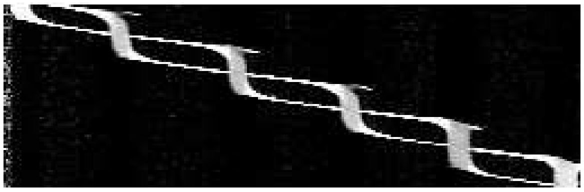

Figure 7: XT-slice near ankle run through a change detection operation.

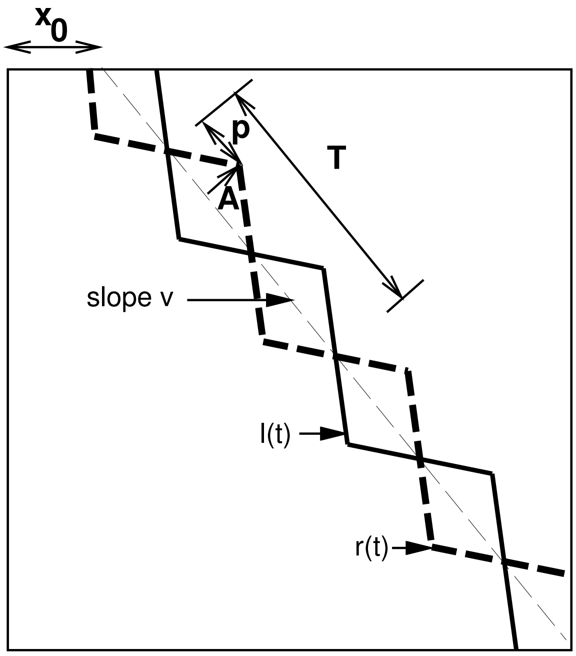

Figure 8: Template model used on ankle XT-slice. Template has three varying parameters (A, T and p) and two fixed parameters (x_0 and v).

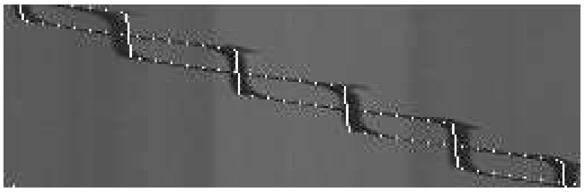

Figure 9: Ankle XT-slice, with template match signals overlaid in white. Note that there are two signals $l(t)$ and $r(t)$ which define each template.

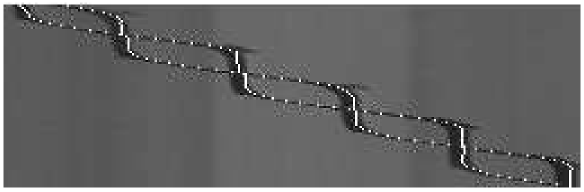

Figure 10: Using the template match as a initial position for the snake, snakes refine the estimate. This yields the centers of both ankles as they move over time. Resulting snakes are shown in white.

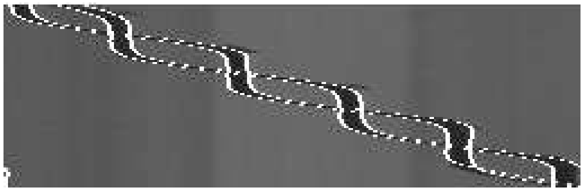

Figure 11: By attracting the snakes to the positive and negative spatial derivatives of the change detection output, we obtain the four spatiotemporal edges corresponding to the front and back of each leg.

such an energy function for both signals results in a snake fit as shown in Figure 10. In doing so, we refine our estimate of the centers of both ankles.

However, we wish to recover the *bounding contours* of the body. Each snake is split up into two new snakes, for a total of four snakes. Two of them are designed to track the left leg; two of them track the right. This is achievable by splitting the two snakes into four and attracting snakes to the positive and negative blurred spatial derivative of the change detection algorithm output, i.e. if $l(t)$ and $r(t)$ represent the results of the template match on image $O(x,t)$, then $l_l(t) = l(t)$ and $r_l(t) = r(t)$ use an energy function $gaussian(x)*derivative(x)*O(x,t)$ and $r_l(t) = l(t)$ and $r_r(t) = r(t)$ use an energy function $-gaussian(x)*derivative(x)*O(x,t)$. Figure 11 shows the four recovered spatiotemporal edges.

Body contour following. Since the body is spatially contiguous, the XT-slice at one height is very similar to XT-slices at nearby heights. It follows that the spatiotemporal edges at one XT-slice are similar to spatiotemporal edges at another. Thus the spatiotemporal edges recovered at one height can be used as an initial configuration for snakes at a nearby XT-slice. In this fashion the whole body contour can be recovered, from head to toe. So, we recover four spatiotemporal edges as a function of height (y), represented by $l_l(y,t)$, $l_r(y,t)$, $r_l(y,t)$, and $r_r(y,t)$. Tracked body contours are shown in Figure 12a. Near the hip, two pairs of snakes should merge ideally, so computing four snakes above torso is not necessary.

Upsampling body contours. All the processing described thus far was done on a low-resolution image sequence, obtained by downsampling the original image sequence twice. It is straightforward to upsample the contours recovered at a lower scale; assuming a factor of k between scales, we obtain four new spatiotemporal edges at each height with: $l_{l2}(y,t) = kl_l(\frac{y}{k},t)$, $r_{l2}(t) = kr_{l2}(\frac{y}{k},t)$; $r_{l2}(t) = kr_l(\frac{y}{k},t)$ $r_{r2}(t) = kr_r(\frac{y}{k},t)$. The upsampled contours give an estimate of the contour at each y location at a higher resolution. These are refined, just as before, with snakes from the height y of the head to the height y at the toe. Each upsampled contour is used as an the initial configuration for a snake; the image energy function used is merely the positive (for $l_{l2}(y,t)$ and $r_{l2}(y,t)$) and negative (for $l_{r2}(y,t)$ and $r_{r2}(y,t)$) blurred spatial derivative of the change detection algorithm output. Since we have estimates for each XT-slice from a lower resolution, processing can occur in parallel for each y location. The results of processing on higher scales are shown in Figure 12b,c.

5 Gait modeling

Figure 13 shows the body contours overlaid on one of the middle frames of the image sequence. In order to recover exact locations of the head, the hip joint, the two knee joints, and two ankle joints accurately in each frame, more processing is required. A variety of algorithms can be imagined that work from these contours.

The one that we use is as follows. Average the four left and right body contours to form two "skeletons." Figure 14 shows one skeleton overlaid on one frame from the image sequence. To recover angle signals, we perform line-fit operations on the skeleton at appropriate locations. Which locations should we use? Since we know roughly where the walkers head and toe are, and since humans have knees, hips and ankles at predictable locations, we can perform line fits between heights where the hip probably is and where the knee probably is to recover upper leg angle information, and likewise for the lower leg angle information. This recovery is less accurate when the hips, knees and ankles are not where we expect them to be. Our solution is to do *another* snake operation in XY for each frame in the image sequence, using change detection outputs as energy functions again. Such an energy function encourages the snakes to climb to the middle of the body. We know that there is a a *second-order discontinuity* at the hip, knee, and ankle. Since we have the flexibility in the snake algorithm to set first and second order discontinuities[1], we can set the spline to be second-order discontinuous at those points using the coarse head and toe locations and a simple height model of a human. The snake points are free to move in x and y, so the bends of the hips, knees, and

[1] These are *alpha* and *beta* in the original paper.

Figure 13: One frame from a body-tracked image sequence, with the four recovered body contours for this particular frame overlaid in white.

Figure 14: The four body contours are averaged into two skeletons, shown overlaid in white.

ankles are free to move about. No part of the snake should be outside the walker's body. Figure 15 shows how the snake fit changes. We may wish to stop with these six coordinates, obtained by processing each frame: (x_{head}, y_{head}), (x_{hip}, y_{hip}), (x_{knee1}, y_{knee1}) and (x_{ankle1}, y_{ankle1}), (x_{knee2}, y_{knee2}) and (x_{ankle2}, y_{ankle2}). However, we found that using these six coordinates to do line-fits yields much better data than just using these six coordinates. Figure 16 shows the recovered stick model for a particular frame.

6 Gait recognition

The stick model recovered is merely four angle signals that change as a function of time. Upper leg and lower leg angle signals recovered from one image sequence are shown in Figure 17. As expected, the signals are roughly periodic, and left and right leg signals are out of phase.

For recognition of gait patterns at different

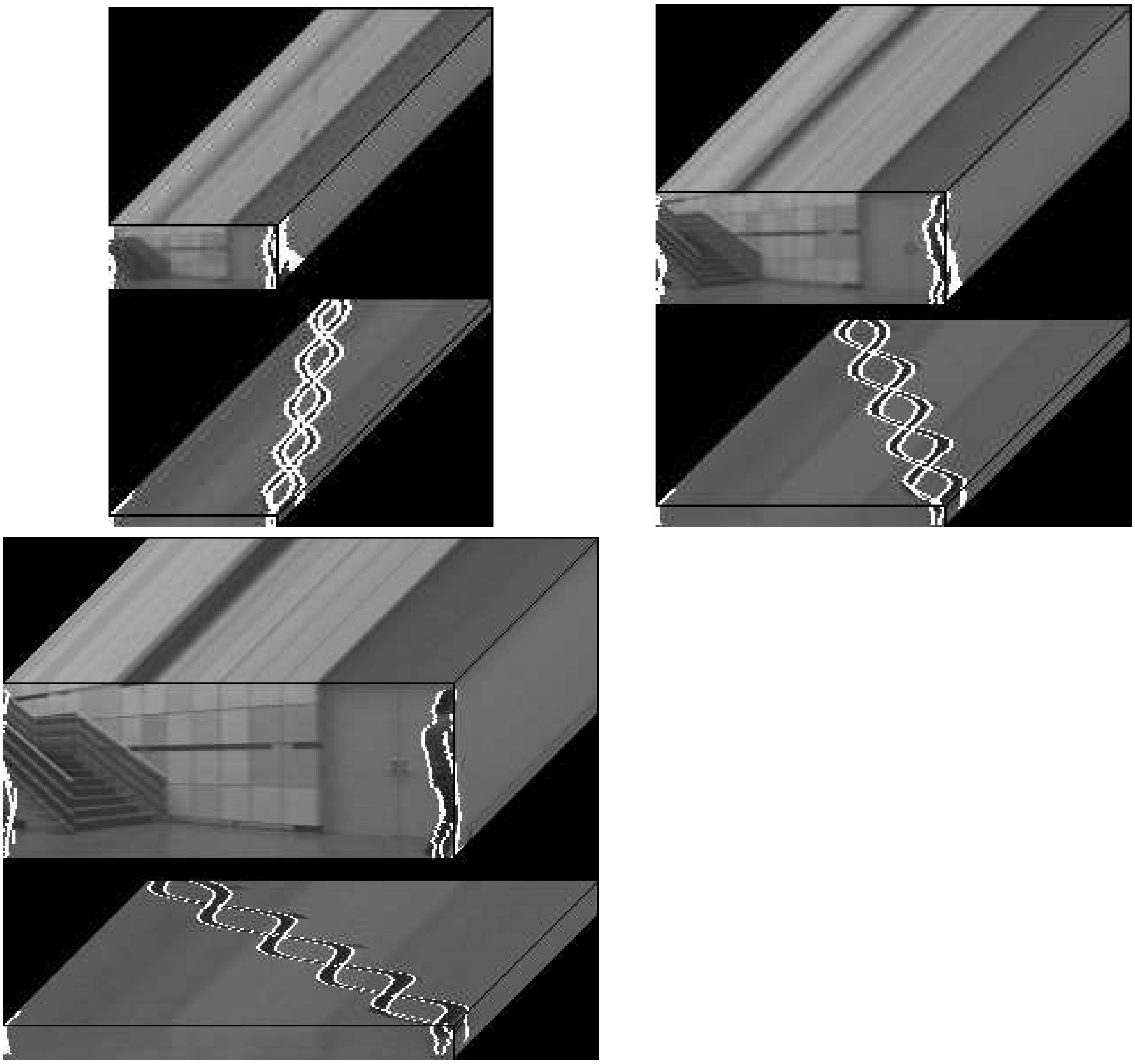

Figure 12: Body tracking at all scales. Snake outputs at all XT-slices and all y values are upsampled and used as initial snake values at higher resolution. (a) Upper left: Body contours recovered for a low resolution image sequence obtained by recovering spatiotemporal edges of each XT-slice. (b) Upper right: body contours recovered for an image sequence at a higher resolution, obtained by running snake algorithm on upsampled contours from lower resolution image sequence. (c) Bottom: Body contours for image sequence at highest resolution.

Figure 15: The skeleton model is refined using snakes, with second order discontinuities inserted at the hip, knee, and ankle. The refined skeleton for a particular time instant is shown in white.

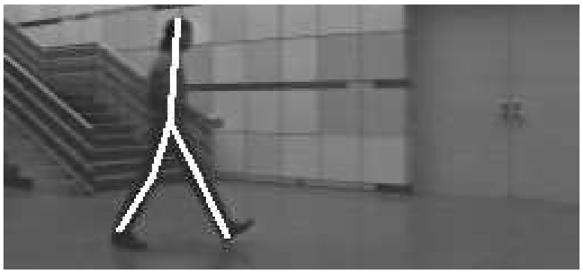

Figure 16: Simple line fitting operations generate a five-stick model. The recovered stick model for a particular time instant is overlaid in white.

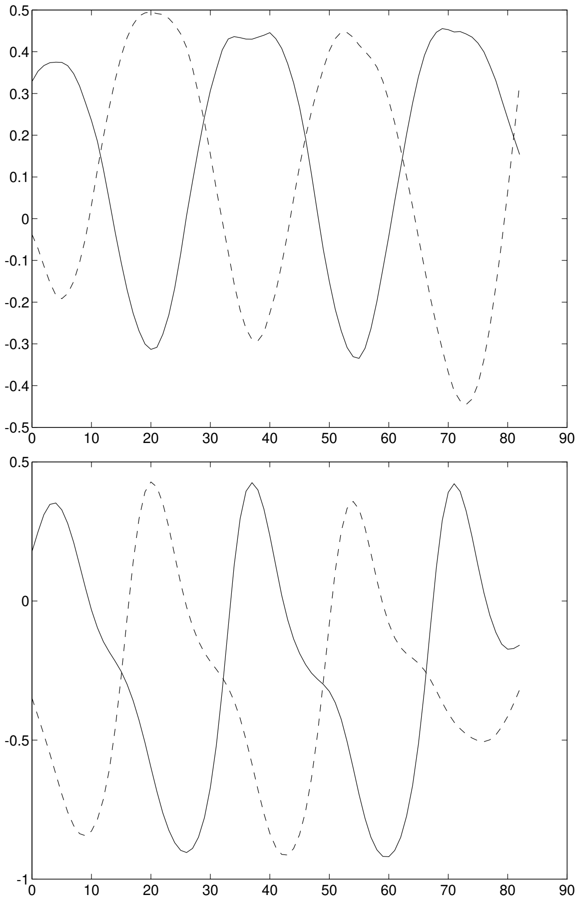

Figure 17: Upper (shown at top) and lower (shown at bottom) leg angle signals for both legs recovered for one walker. (Left leg - solid, Right leg - dashed)

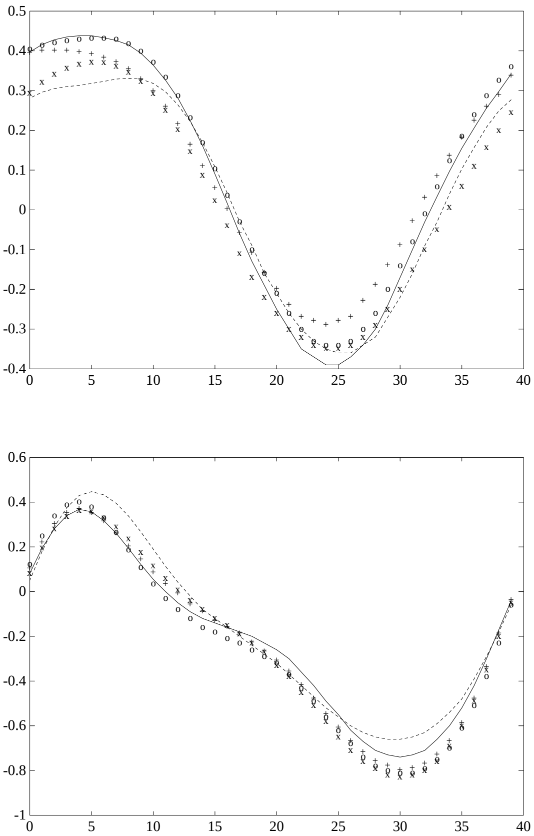

Figure 18: Upper (shown at top) and lower (shown at bottom) leg gait signals (40-dimensional vector) for five walkers. (SI - solid, SAN - dashed, RWP - plusses, AJA - circles, LWC - Xs)

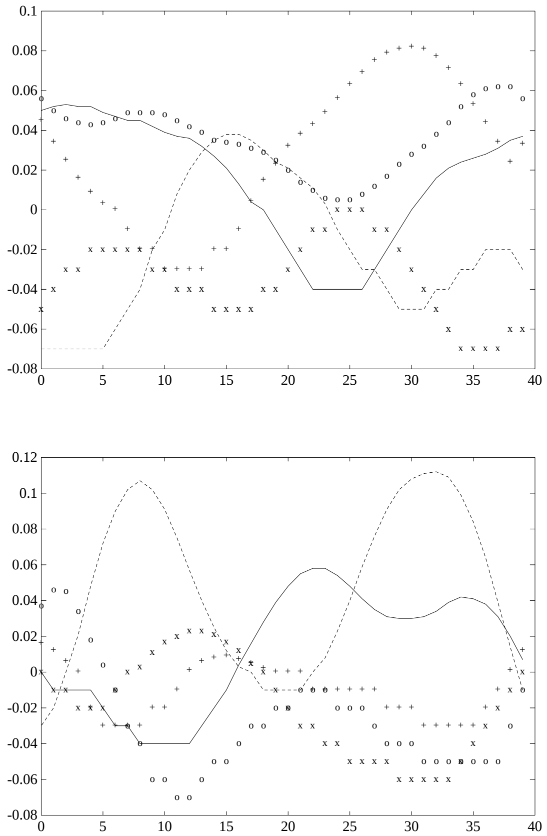

Figure 19: Deviation from mean upper (shown at top) and lower (shown at bottom) leg gait signals for the same five walkers. (SI - solid, SAN - dashed, RWP - plusses, AJA - circles, LWC - Xs)

walking speeds, some time-warping of these four walking signals is necessary. Data used for recognition were extracted by: (a) finding zero-crossings of the derivative of one of the lower leg angle signals; (b) choosing only the zero-crossings with a negative derivative, and using these as indices to copy signal data of both the lower leg signal and the upper leg signal into two vectors; (c) linearly interpolating these two vectors, and joining them to a produce a fixed length vector; (d) walkers walking from right to left had this vector multiplied by -1.

We ran the entire algorithm on 24 different image sequences containing frontoparallel walkers. Four were of AJA, six were of LWC, four were of RWP, six of SAN, and six of SI. Image sequences were taken indoors at three different locations at three different times in the day, separated by roughly an hour. Averaged upper and lower leg vectors for the five walkers (i.e. averaging all the vectors obtained from a particular walker into one vector) are shown in Figure 18. To see the features more clearly, refer to Figure 19 to observe how gait signals deviate from the mean gait signal of all individuals.

Using a simple recognition technique, k-nearest neighbors with Euclidean distance measures, worked reasonably well. It runs as follows. To classify a particular image sequence with extracted gait vectors v_1, v_2, ... v_m with a table of previously classified walks w_1, ... w_n, classify each v_i independently, and classify the image sequence as belonging to the class chosen the most often out of all the vectors v_i. Each vector v_i is classified by computing a Euclidean distance $D_{ij} = ||v_i - w_j||^2$ between v_i and each walk w_j in the dictionary; the most common class in the k smallest distances is chosen as the classification of the vector.

Nineteen of the twenty four image sequences were correctly recognized with with $k = 5$, and 17 or 18 were recognized with $k = 3, 4, 6$. Since chance recognition rate is 20%, a recognition rate of 79% is promising. It is difficult to make a scalability claim of the recognition rates without accumulating and measuring more data. Naturally, with more measurements of each walker's walking pattern, accuracy will increase; with more walkers to choose among, however, accuracy will decrease. However, it is unrealistic to expect near-perfect gait recognition performance. Instead, gait recognition will be most promising when combined with other recognition techniques.

7 Conclusion

We show a method for recovering a stick model of a human by spatiotemporal analysis of gait patterns. The initial model of the walker is simple; a walker is a translating blob which has braided spatiotemporal patterns in the lower half of his body. By recognizing these spatiotemporal signatures, we can impose a model for subsequent spatiotemporal analysis. This allows us to recover the spatiotemporal edges of the walker. This recovery process yields promising results in a new recognition problem.

8 Acknowledgments

This work was sponsored by Goldstar and TVOT. Image sequences were recorded and digitized with the aid of Stephen Intille and Lee Campbell. Thanks go to Lee Campbell, Stephen Intille, Ali Azarbayejani and Rosalind Picard for posing as walkers.

References

[1] Chen, Z. and Lee, H. Knowledge-guided visual perception of 3-d human gait from a single image sequence. *IEEE Transactions on Systems, Man, and Cybernetics* 22:336-342.

[2] Cipolla, R. and Yamamoto, M. Stereoscopic tracking of bodies in motion. *Image and Vision Computing* 8: 85-90, 1990

[3] Cutting, J.E. and Kozlowski, L. Recognizing friend by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society* 9:353-356, 1977.

[4] Goddard, N. The perception of articulated motion: Recognizing moving light displays, University of Rochester Ph.D. thesis, June 1992.

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

[6] Kurakake, S. and Nevatia, R. Description and tracking of moving articulated objects, *ICPR* pp. 491-495, 1992

[7] Leung, M. and Yang, Y.H. Human body motion segmentation in a complex scene. *Pattern Recognition* 20: 55-64, 1987.

[8] Leung, M. and Yang, Y.H. A region based approach for human body motion analysis. *Pattern Recognition* 20: 321-339, 1987.

[9] Kass, M., Witkin, A. and Terzopoulos, D. SNAKES: Active contour models. *Intern. J. Computer Vision* 1: 321-332.

[10] Kozlowski, L. and Cutting, J. Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychophysics*, 21: 575-580, 1977.

[11] Marr, D. and Nishihara, H.K. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. London B* 200: 269-294, 1978.

[12] O'Rourke, J. and Badler, N. Model-based image analysis of human motion using constraint propagation. *IEEE PAMI* 2: 522-536.

[13] Rashid, R. Towards a system for the interpretation of moving light displays. *IEEE PAMI* 2: 574-581, 1980.

[14] Polana, R. and Nelson, R. Detecting activities. *CVPR* 2-7, 1993.

[15] Qian, R.J. and Huang, T.S. Motion analysis of human ambulatory patterns, *ICPR* 220-223, 1992.

[16] Rohr, K. Incremental recognition of pedestrians from image sequences. *CVPR* pp. 8-13, 1993.

[17] Shio, A. and Sklansky, J. Segmentation of people in motion. IEEE Workshop on Visual Motion, pp. 325-332, 1991.

[18] Tsukiyama, T. and Shirai, Y. Detection of the movements of persons from a sparse sequence of TV images. *Pattern Recognition* 18: 207-213, 1985.

[19] Webb, J.A. and Agarwal, J.K. Structure from motion of rigid and jointed objects. *Artificial Intelligence*, 19: 107-130, 1982

[20] Yamamoto, M. and Koshikawa, K. Human motion analysis based on a robot arm model, *CVPR*, pp. 664-665, 1991.