This is a preprint of an article published in Computer Animation and Virtual Worlds, **15**(3-4):399-406, 2004.

This journal may be found at:
http://www.interscience.wiley.com

Automated Markerless Extraction of Walking People Using Deformable Contour Models

David K Wagg and Mark S Nixon
School of Electronics and Computer Science
University of Southampton, SO17 1BJ, UK
{dkw02r, msn}@ecs.soton.ac.uk

Abstract

We develop a new automated markerless motion capture system for the analysis of walking people. We employ global evidence gathering techniques guided by biomechanical analysis to robustly extract articulated motion. This forms a basis for new deformable contour models, using local image cues to capture shape and motion at a more detailed level. We extend the greedy snake formulation to include temporal constraints and occlusion modelling, increasing the capability of this technique when dealing with cluttered and self-occluding extraction targets.

This approach is evaluated on a large database of indoor and outdoor video data, demonstrating fast and autonomous motion capture for walking people.

KEY WORDS: motion capture, gait, walking, snakes

Introduction

In recent years, interest in human motion analysis has increased rapidly, due mostly to the large number of potential applications for this technology [1, 2]. Fields such as computer animation, film and TV production, model-based video coding and smart surveillance would benefit immensely from an improved ability to automatically extract human motion from video data. However, the variability in appearance, and range of motion possible in typical human activities, make this problem very difficult to solve.

Current motion capture systems operate at the cost of attaching markers to the subject. These markers are then tracked using optical or electro-magnetic sensing systems, avoiding the issues involved in tracking people themselves. While successful in some applications, marker-based systems suffer from a number of disadvantages which limit their deployment. The most obvious is their often prohibitive cost, but additionally, the markers employed can restrict the range of motion of the subject. The cooperation of the subject is also required, making such systems useless for surveillance applications.

These limitations motivate the development of markerless motion capture systems. Current approaches typically fall into one of two categories; those that attempt to recover general, full-body motion, and those that focus on specific, limited activities.

Recovering full body motion is naturally more difficult, due to the increased range of possible motions and the greater incidence of self-occlusion. Many recent approaches to this problem have employed multiple cameras [3, 4, 5, 6] to resolve pose ambiguities. However, this approach is more expensive in monetary terms and computational complexity, and is still unsuitable for some applications. No markerless system has yet demonstrated fast, reliable capture of unconstrained full-body motion.

Although this capability is our eventual goal, it can be beneficial to solve more constrained problems first, applying the techniques learned to unconstrained motion. Most research to date following this approach has focused on people walking and running [7, 8, 9, 10], as these activities account for the majority of everyday human motion.

Regardless of approach, almost without exception recent approaches have utilised some form of anatomical shape model to aid the motion capture process. For constrained motion, it is often possible to apply models of motion as well. Model-based approaches incorporate knowledge of the shape and dynamics of human motion into the extraction process, ensuring that only image data corresponding to allowable human shape and motion is extracted. However, models present their own problems. The more accurate a model is required to be, the greater the number of parameters are required to define the model. In general, computational complexity increases exponentially with model complexity.

This paper extends research presented in [11] to automatically extract walking persons for the purpose of identification by gait. The model-based methodology presented is fast, robust and completely automated. However, computational concerns limit the models used to simple geometric shapes with low degrees of freedom. We work round this problem by using shape and

motion models to initialise a deformable contour model. We use the model prediction as an approximation of the actual data, and allow small deviations from the model based on local image cues.

We assume that a single subject is present in the scene, moving at an approximately constant speed parallel to the camera view plane, against a cluttered background. Although these are relatively constrained capture conditions, our extraction method is capable of operating autonomously, in outdoor conditions with high degrees of clutter.

Global motion of the subject is determined by temporal accumulation, applying anatomical constraints in a hierarchical fashion to extract body shape parameters. We establish gait period and phase independently of shape parameters, limiting computational demands and possible initialisation errors. This motion information is combined with prior knowledge of mean joint motion during normal gait, to create an approximate model of the subject's leg motion. Finally we use this model to initialise a deformable contour model, allowing local adaptation to fit the contours to observed data. We present results of the motion capture process on clean data filmed within the laboratory, and on real-world data, showing reasonable extraction performance even in adverse conditions.

Global Motion and Shape Extraction

For walking people, motion is dominated by velocity in the horizontal plane. This naturally motivates a hierarchical decimation of motion, determining first the largest motion components and subsequently smaller motion components (Fig. 1).

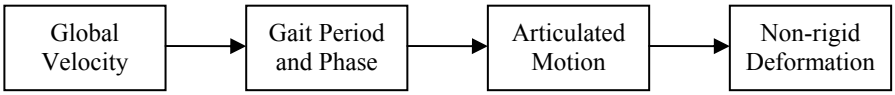


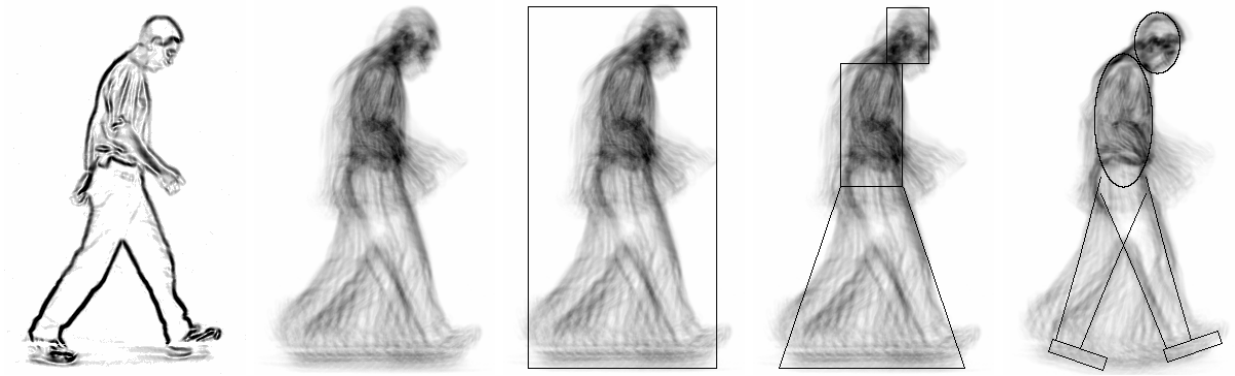
Figure 1: Motion Extraction Hierarchy

Our first step is to apply edge detection and background subtraction to the raw video data, in order to reduce its inherent redundancy (Fig. 2a). For simplicity, our background model is computed as a temporal median of neighbouring frames, although more sophisticated modelling strategies are possible. This pre-processing simplifies analysis by reducing the scene to moving edges only, including the person of interest.

We may determine horizontal motion independently of shape parameters by temporally accumulating edge data according to expected velocity [11]:

$$A_v(i,j) = \sum_{t=0}^{N-1} E_t\left(i + v\left(\frac{N}{2} - t\right), j - dy_t\right) \tag{1}$$

Where A_v is the accumulation for velocity v (in pixels per frame), E_t is the edge strength image at frame t, i and j are coordinate indices, N is the number of frames in the gait sequence and dy_t defines the vertical oscillation of the subject. This vertical motion is initially unknown and set to zero. However, after determining gait period and phase this motion can be estimated (Eqn. 3), permitting an improved temporal accumulation (Fig. 2e). At the correct accumulation velocity for a moving object, its edges at each frame will accumulate to a coherent global average view (Fig. 2b).



(a) Section of pre-processed image data | (b) Global temporal accumulation | (c) Bounding box fitted | (d) Approximate shape estimate | (e) Rigid articulated model

Figure 2: Global tracking and shape estimation from temporal accumulation

Each moving object in the scene will appear as a peak in a plot of maximal accumulation intensity against velocity. If the subject is the most significant moving object in the scene (in terms of edge strength and visibility), their velocity can be inferred by selecting the highest peak in this plot. However, for outdoor imagery this assumption is often violated. In these cases, we can use the expected shape of the subject to disambiguate them from other moving objects (Figs. 2c, 2d). This shape expectation is computed *a priori* from mean anatomical data [12] scaled to the subject's apparent height (as measured from their bounding box). The optimal subject velocity is then the velocity that maximises correlation of the subject's shape template (Fig. 2d) with the accumulated edge intensity image (A_v).

By this process we derive an initial estimate of the subject's starting position, velocity and size, sufficient for estimation of their gait period. We can further improve our shape model by template matching ellipses against the temporal accumulation (Fig. 2e), thus estimating coarse shape in a robust fashion. Leg parameters are initially set to fixed proportions of the person's height.

Global Articulated Motion Extraction

The motion of the leg during normal gait is periodic, and this motion may be estimated by general methods for periodicity detection, avoiding the use of complex models. Motion periodicity is determined by measuring some quantity related to shape over time and analysing this signal for periodicity. Cutler *et al*. [13] present a general method for periodicity detection by measuring silhouette self-similarity over time, using autocorrelation-based analysis to extract the gait period. However, this method has relatively high computational demands, particularly for long video sequences. Other common methods involve analysing periodicity in silhouette width or height [14], which result in far lower computational requirements. We employ a similar strategy [11], measuring instead the sum of edge intensity within the outer region of the subject's legs throughout the gait sequence (Fig. 3).

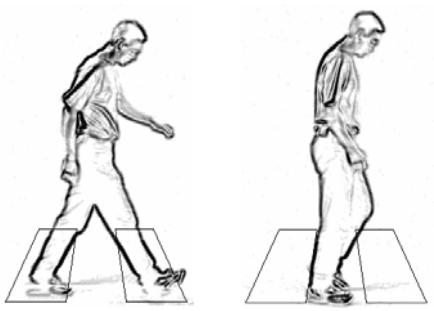
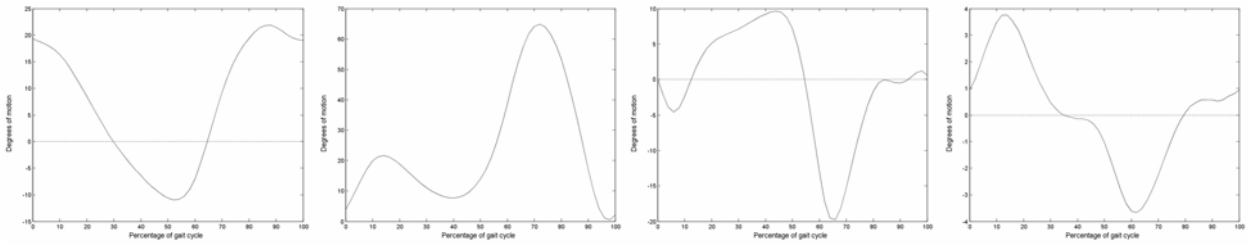


Figure 3: Gait period estimation

We use data collected from clinical gait studies [15, 16] to build prototypical models for hip, knee, ankle and pelvis rotation. Fig. 4 shows these mean rotation models for a single gait cycle, from right heel-strike to right heel-strike.



(a) Hip rotation (b) Knee rotation (c) Ankle rotation (d) Pelvic list

Figure 4: Mean joint rotation patterns

Movement of the pelvis includes both axial rotation and list (resulting in horizontal and vertical oscillation respectively in the sagittal plane), but pelvic axial rotation is simple enough to be modelled by a single sinusoid, with a minimal degree of error:

$$\theta_p(t) = A_p \cos(wt + \phi_h) \tag{2}$$

Where θ_p is pelvic axial rotation, A_p is the amplitude of rotation (approximately 5° for normal gait), w is the gait frequency and ϕ_h is the starting gait phase. Although the magnitude of pelvic rotation is small, accounting for this source of variation in hip joint position can significantly reduce errors in the estimation of hip joint rotation.

The vertical oscillation of the subject's upper body is also modelled by a single sinusoid, with parameters in fixed proportion to the subject's height and gait motion:

$$dy_t = A_y \sin 2\left(wt + \phi_h + \pi/8\right) \tag{3}$$

Where dy_t is the y-displacement of the centre of the torso at frame t (see Eqn. 1), A_y is the amplitude of oscillation, w is the gait frequency and ϕ_h is the starting gait phase.

Using our estimates of the subject's gait frequency and starting phase these models are scaled to fit the subject, using Hermite spline interpolation. This yields an initial estimate of the subject's leg motion, providing a good basis for local

adaptation. Arm motion is not included in our models at the present time, although it should be possible in the future to include it in a similar manner as for leg motion.

The use of a parametric model of motion increases the robustness of motion capture through multiple averaging processes. However, this averaging also decreases accuracy. We effectively assume identical motion for each leg and identical motion across gait cycles, which causes mismatch because motion will vary under normal conditions. There is also mismatch due to individual variation in gait patterns (which forms the basis for recognition by gait). It is possible to retain these assumptions and adapt these models to better fit the observed data [17]; however, rigid models will never match observed data exactly without an implausibly large number of parameters. In this work we attempt to solve this problem through the use of local adaptation processes. The rigid model extraction is used as a robust, though not highly accurate, initialisation for a deformable contour model. We can then allow the contours to deform according to local image cues, under spatio-temporal continuity constraints.

Global Leg Shape Extraction

Before initialising our deformable contours, it is desirable to improve the shape of our model, reducing the amount of (less robust) local adaptation required. Using a global estimation technique means that shape can be estimated from the whole sequence of images of the person, avoiding errors due to localised noise or occlusions.

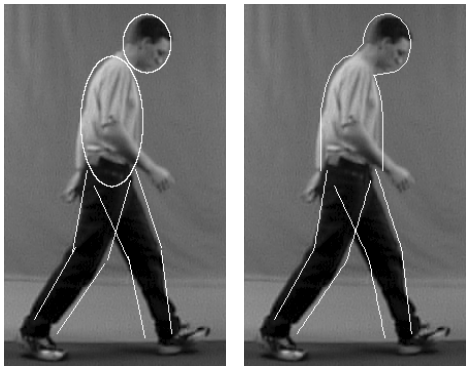
This re-estimation process is necessary because our initial estimate of leg shape based on the height of the person may not be appropriate for certain types of clothing (baggy trousers, shorts or skirts for example). An improved estimate is obtained by computing a line Hough transform for each frame within the upper and lower leg regions (above and below knee level). Within each Hough space we find the pair of accumulation peaks satisfying constraints on the expected line orientation and the distance between the two lines (leg width), yielding an estimate of leg shape for that frame. Final estimates of leg width are computed as the mean of the best parameters from each frame, weighted by accumulation intensity:

$$w_m = \sum_{t=0}^{N-1} p_t w_t \Big/ \sum_{t=0}^{N-1} p_t \tag{4}$$

Where w_m is the mean width of the leg (at the hip, knee or ankle) over N frames, and w_t and p_t are the estimated leg width and the peak accumulation intensity respectively at frame t. This process yields estimates of width at the hip and ankle, and two at the knee (from upper leg and lower leg estimation), allowing for discontinuity in leg width at the knee (caused by a skirt or shorts).

Local Model Deformation

We use our initial rigid model fit as the starting point for a deformable contour model. To adapt the contour shape to fit the image data, we employ a relatively simple gradient descent formulation based on the greedy snake [18], whereby contour shape adaptation is expressed as a process of energy minimisation. Snake energy incorporates internal constraints on local curvature and contour point spacing, and external (image) constraints used to attract the snake to image features (edges in our case):



(a) Rigid geometric model (b) Deformable contour model

Figure 5: Contour model initialisation

$$E^*_{snake}[v(s)] = \min \int_0^1 \left(E_{\text{int}}[v(s)] + E_{ext}[v(s)]\right) ds \tag{5}$$

Where E^*_{snake} is the minimal snake energy, $\boldsymbol{v(s)}$ is the snake contour, E_{int} is the internal snake energy, and E_{ext} is the external energy. Internal energy for all contours is described by:

$$E_{int}[v(s)] = \alpha E_{cont} + \beta E_{curv} + \gamma E_{temp} \tag{6}$$

Where E_{cont} corresponds to normalised first-order continuity of the snake, E_{curv} corresponds to normalised second-order continuity of the snake and E_{temp} corresponds to normalised first-order temporal continuity. The weighting coefficients α, β and γ control the balance of these three energy contributions.

The upper body and leg contours differ greatly in shape and expected level of occlusion. In order to account for this difference, we use slightly different external energy terms to optimise snake evolution. For the upper body contours the external energy is described by:

$$E_{ext,\ body}[v(s)] = \lambda I_t[v(s)] \tag{7}$$

Where I_t is the image attraction term for the contour (smoothed edge intensity), and λ is a weighting term. For the leg contours we add an occlusion weighting term $o[\mathbf{v}(s), \theta_h]$, and an additional constraint E_{side} forcing the front and back leg contours to remain within an expected distance:

$$E_{ext,leg}[v(s)] = o[v(s), \theta_h]\lambda I_t[v(s)] + \rho E_{side}[v(s)] \tag{8}$$

Where $o[\mathbf{v}(s), \theta_h]$ is the occlusion model prediction for the contour $\mathbf{v}(s)$ at hip phase (pose) θ_h and E_{side} is equal to the difference between the expected and measured distance between the front and back contours of the leg:

$$E_{side}[v(s)] = \left| w_m[v(s)] - \left| v_{front}(s) - v_{back}(s) \right| \right| \tag{9}$$

Where w_m defines the expected leg width at each contour point, v_{front} and v_{back} are the front and back contours of the leg. All snake control parameters were determined empirically; optimal values will vary depending on the specific application.

We compute the occlusion model *a priori*, assuming mean gait motion and leg shape. This model (Fig. 6) defines the expected level of occlusion at each contour point for each leg position during a gait cycle. This is computed by assigning 0 to the model if the point is occluded, or 1 if it is not, followed by an appropriate degree of spatial and temporal smoothing to yield a continuous model. The purpose of the occlusion model is to reduce the contribution of image features at points in the gait cycle where the legs occlude each other. At these points we would not expect to see reliable edge information, and so we force the snake to rely more on initialisation, and on internal and temporal constraints (the temporal constraint effectively allows some degree of interpolation over occluded frames). The snake contours are driven to a minimal energy state by an iterative process of gradient descent. Note that due to the inclusion of a temporal constraint, the order of iteration in performing the minimisation is important. We perform a single iteration of gradient descent for each frame in the sequence, before repeating the process for subsequent iterations.

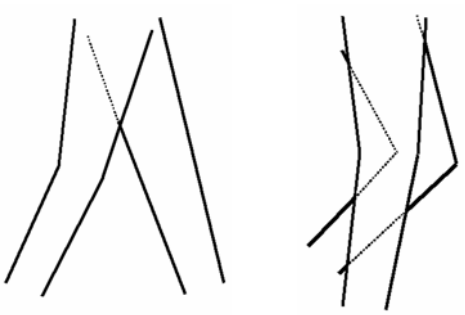


Figure 6: Leg occlusion model

A final point on this adaptation process is that due to its local nature, if the initial contour is too far from the correct solution, it will get stuck in local minima (irrelevant edges). This problem is partially avoided by re-estimating a global (skeletal) motion model from our extracted contours, relying on the averaging process to eliminate poor contour extractions. We then repeat the local adaptation process using this improved initialisation. However, errors are still possible if the initial model is not accurate enough for the observed motion.

Results

The performance of the motion capture process was evaluated on the Southampton HiD database [19]. Each subject was filmed from a fronto-parallel viewpoint, in controlled laboratory conditions and in outdoor conditions. The database is encoded in Digital Video (DV) format at a resolution of 720x576 pixels, recorded at a rate of 25 frames per second with approximately 90 frames per gait sequence. The database includes 2163 indoor and 2661 outdoor gait sequences split over 115 subjects. A 2.4GHz Pentium 4-based PC was used for all testing, requiring approximately 15 hours in pre-processing and 20 hours in gait extraction for the whole database, equivalent to an overall processing rate of approximately 3 frames per second. The system is fully automated, requiring no human intervention to aid the analysis.

Figs. 7-10 show examples of the extraction obtained on indoor and outdoor data:

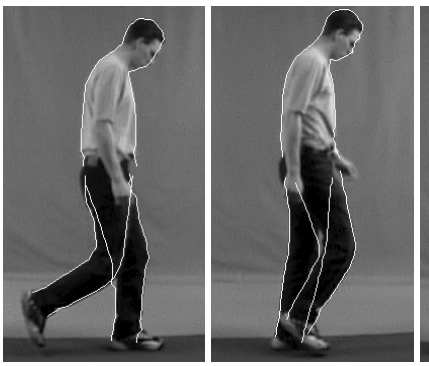


Figure 7: Subject 13 (indoor dataset example)

Figure 8: Subject 13 (outdoor dataset example)

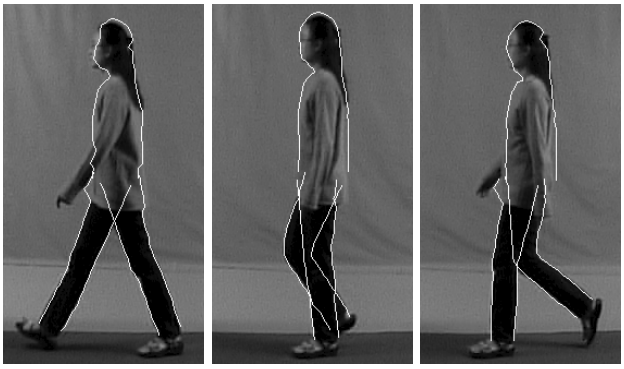


Figure 9: Subject 37 (indoor dataset example)

Figure 10: Subject 37 (outdoor dataset example)

Extraction performance is generally good on the indoor data, with accuracy dropping slightly on the outdoor data due to the decreased reliability of edge cues. This problem may be ameliorated by using colour cues in addition to edges, or by employing more sophisticated background modelling strategies in pre-processing. Self-occlusion of the legs is generally handled well, but there are some artefacts caused by occlusions from the arms. This problem may be mitigated by connecting the leg and body contour end-points, forcing them to converge. Alternatively, arm motion could be incorporated into the occlusion model, applying the same principles used in modelling the motion of the legs.

In order to generalise our analysis to the whole database, we measure the variation of mean leg shape over all sequences of each subject:

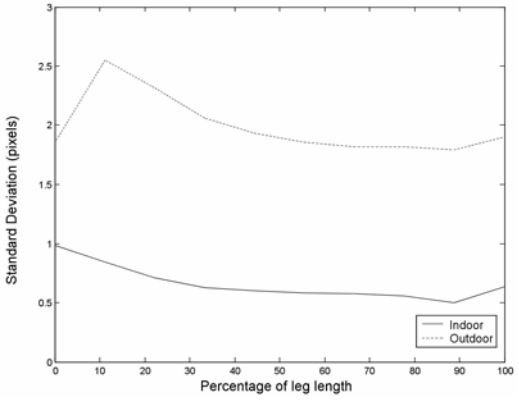


Figure 11: Leg shape consistency

$$\sigma[i] = \sqrt{\sum_{n}^{N-1} \frac{(x_n[i] - \mu[i])^2}{N-1}} \tag{10}$$

Where σ is the standard deviation of mean leg width, i is a leg length index, x_n is the mean leg width measured over sequence n, N is the number of sequences featuring the subject and μ is the mean leg width measured over all sequences of the subject.

This analysis allows us to make comparisons between the indoor and outdoor datasets. For robust extraction performance we would expect a similar extracted leg shape for each sequence, and so variance should be low. High variance in extracted leg shape would suggest inconsistencies in extraction on some sequences. Fig. 11 depicts the average leg shape variation for each dataset, as a function of leg length (0% at the hip to 100% at the ankle). This analysis reinforces the conclusions drawn from a visual examination of extraction performance, showing that leg shape is generally less reliable at the hip (where the arms occlude the legs), and at the ankle (where the foot joins the leg). The reduced reliability of motion capture on the outdoor database is also clear from this analysis.

Conclusions

We have presented a new fully automated model-based method of motion capture for walking persons, based on the hierarchical application of rigid models and locally deformable models. This yields fast and reasonably accurate operation, and has proven capable of handling a large database of indoor and outdoor data without human intervention. Mixing rigid models with deformable models goes some way to alleviate the conflict between accuracy of extraction and model complexity. However, our current approach separating the two models may not be ideal; performance may be improved by combining the two processes.

Although there are some issues concerning the reliability of edge cues in noisy outdoor data, extraction performance at this early stage is encouraging. Future research will focus on improving performance on outdoor data, aiming to reduce the need to perform motion capture in restrictive studio environments.

Acknowledgements

This research is partially funded by the European Research Office of the US Army under Contract No. N68171-01-C-9002. David Wagg gratefully acknowledges support by the UK Engineering and Physical Sciences Research Council (EPSRC).

References

[1] T B Moeslund and E Granum. "A Survey of Computer Vision-Based Human Motion Capture." *Computer Vision and Image Understanding*, **81** (3), pp. 231-268, 2001.

[2] L Wang, W Hu and T Tan. "Recent Developments in Human Motion Analysis." *Pattern Recognition*, **36** (3), pp. 585-601, 2003.

[3] D Gavrila and L Davis. "3-D model-based tracking of humans in action: a multi-view approach." *Proc. Computer Vision and Pattern Recognition*, pp. 73-80, 1996.

[4] G K M Cheung, S Baker and T Kanade. "Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture." *Proc. Computer Vision and Pattern Recognition*, pp. 77-84, 2003.

[5] A J Davison, J Deutscher and I D Reid. "Markerless Motion Capture of Complex Full-Body Movement for Character Animation." *Proc. Computer Animation and Simulation*, pp. 3-14, 2001.

[6] R Plänkers and P Fua. "Articulated Soft Objects for Multi-View Shape and Motion Capture." *Proc. International Conference on Computer Vision*, pp. 394-401, 2003.

[7] I Haritaoglu, D Harwood, and L S Davis. "W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People." *Proc. Face and Gesture Recognition*, pp. 222-227, 1998.

[8] H Ning, L Wang, W Hu and T Tan. "Articulated Model-Based People Tracking Using Motion Models." *Proc. Int. Conf. on Multimodal Interfaces*, pp. 383-388, 2002.

[9] S Wachter and H H Nagel. "Tracking Persons in Monocular Image Sequences." *Computer Vision and Image Understanding*, **74** (3), pp. 174-192, 1999.

[10] C Yam, M S Nixon and J N Carter. "On the Relationship of Human Walking and Running: Automatic Person Identification by Gait." *Proc. International Conference on Pattern Recognition*, pp. 287-290, 2002.

[11] D K Wagg and M S Nixon. "Model-Based Gait Enrolment in Real-World Imagery." *Proc. Multimodal User Authentication*, pp. 189-195, 2003.

[12] D A Winter. "Biomechanics and Motor Control of Human Movement (2nd Edition)." *John Wiley and Sons*, 1990.

[13] R Cutler and L Davis. "Robust Real-Time Periodic Motion Detection, Analysis, and Applications." *Pattern Analysis and Machine Intelligence*, **22** (8), pp. 781-796, 2000.

[14] C BenAbdelkader, R Cutler and L Davis. "Stride and Cadence as a Biometric in Automatic Person Identification and Verification." *Proc. Face and Gesture Recognition*, pp. 372-377, 2002.

[15] D A Winter. "The Biomechanics and Motor Control of Human Gait: Normal, Elderly and Pathological." *University of Waterloo press, Ontario*, 1991.

[16] M W Whittle and D Levine. "Three-dimensional Relationships between the Movements of the Pelvis and Lumbar Spine during Normal Gait." *Human Movement Science*, **18** (5), pp. 681-692, 1999.

[17] D K Wagg and M S Nixon. "On Automated Model-Based Gait Extraction and Analysis." *Proc. Face and Gesture Recognition*, accepted for publication, May 2004.

[18] D J Williams and M Shah. "A Fast Algorithm for Active Contours and Curvature Estimation." *Computer Vision, Graphics and Image Processing: Image Understanding*, **55** (1), pp. 14-26, 1992.

[19] J D Shutler, M G Grant, M S Nixon and J N Carter. "On a Large Sequence-based Human Gait Database." *Proc. Recent Advances in Soft Computing*, pp. 66-71, 2000.