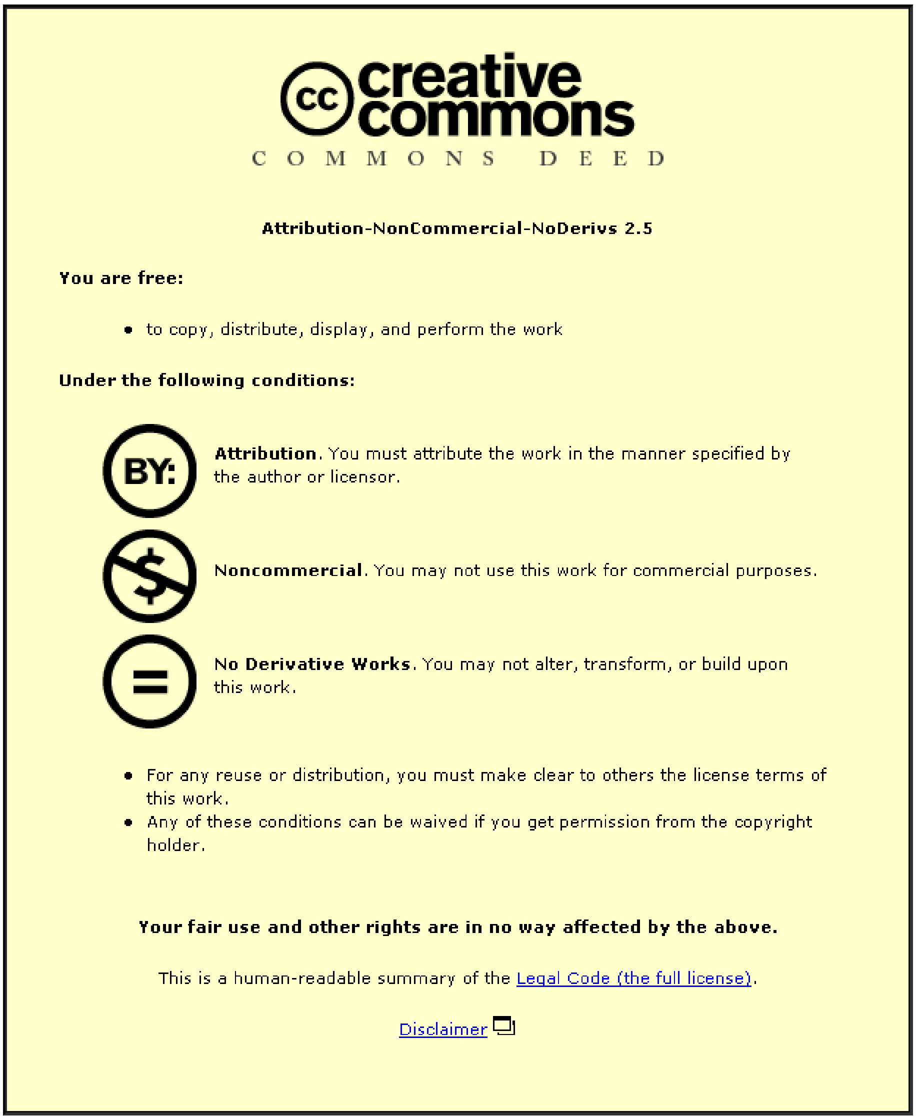


Sports Engineering 7, 53-63, 2004

TECHNICAL NOTE

Model-based automatic tracking of articulated human movement

Maurice R. Yeadon[1], Grant Trewartha[2] and Jon P. Knight[1]

[1] School of Sport and Exercise Sciences, Loughborough University, United Kingdom.
[2] Department of Sport and Exercise Science, University of Bath, United Kingdom.

ABSTRACT

This study applied a vision-based tracking approach to the analysis of articulated, three-dimensional (3D) whole-body human movements. A 3D computer graphics model of the human body was constructed from ellipsoid solids and customised to two gymnasts for size and colour. The model was used in the generation of model images from multiple camera views with simulated environments based on measurements taken on each of three synchronised video cameras and the lighting sources present in the original recording environment. A hierarchical procedure was used whereby the torso was tracked initially to establish whole body position and orientation and subsequently body segments were added successively to the model to establish body configuration. An iterative procedure was used at each stage to optimise each new set of variables using a score based on the RGB colour difference between the model images and video images at each stage. Tracking experiments were carried out on movement sequences using both synthetic and video image data. Promising qualitative results were obtained with consistent model matching in all sequences, including sequences involving whole-body rotational movements. Accurate tracking results were obtained for the synthetic image sequences. Automatic tracking results for the video images were also compared with kinematic estimates obtained via manual digitisation and favourable comparisons were obtained. It is concluded that with further development this model-based approach using colour matching should provide the basis of a robust and accurate tracking system applicable to data collection for biomechanics studies.

Keywords**:** motion capture, marker-free, human modelling, video image

INTRODUCTION

Technological advances have contributed to the increased interest in vision-based/ marker-free approaches to human motion analysis. Such approaches may prove to be a less expensive and more flexible means of human motion capture than commercial opto-electronic systems or laser-based shape measurement systems. However, tracking the kinematics of human motion without artificial markers is a task that is non-trivial due primarily to the complex structure of the human body, the nature of the movements and the regular occurrence of occlusions of one body segment by another. The potential applications of marker-free human movement tracking systems are wide-ranging, including performance analysis, surveillance, and man-machine interfaces including gesture recognition. For application to the collection of kinematic data for biomechanical analyses, the accuracy of the kinematic estimates obtained is of major importance.

The tracking of human movement generally aims to recover the 3D pose (position, orientation and configuration) of the human form over time. This often involves estimating multiple joint angles with respect to an object-centred coordinate system. Previous approaches have demonstrated successful tracking of movements of a constrained nature. For example the movements tracked were essentially 2D or assumed to be limited to specific planes of motion (Yacoob & Davis 2000), the possible model configurations were pre-determined (Rohr 1994), or only simulated sequences were tracked (Chen & Lee 1992). A small number of studies have demonstrated success on less constrained movements with real data using image properties such as: edges (Gavrila & Davis 1996), ellipsoid regions (Bregler & Malik 1998), texture (Lerasle *et al.* 1999) and silhouette (Delamarre & Faugeras 2001). The latter approach demonstrated promising tracking results on running movements although some movements involving longitudinal rotations proved problematic to track.

Despite the large amount of research activity in this area, successful tracking results on general 3D movement from video are still limited (Gavrila 1999). Common threads among the more successful systems include a 3D model-oriented approach and the use of image information from multiple synchronised camera views. Both of these factors reduce the ambiguity of matching 2D features when describing 3D motion. A model-based tracking approach based on the matching of colour pixel values has been applied successfully to the tracking of 3D rigid body motion from synthetic and video data (Trewartha *et al.* 2003). Since human movement may be regarded as the motion of a hierarchical system of linked rigid bodies, it may be possible to extend this method by successively adding body segments after a rigid body tracking of the torso. The purpose of this study is to apply a model-based tracking approach in order to recover motion parameters from a number of synthetic and video sequences containing 3D human body movements. The performance of the tracking procedure will be evaluated by comparing the tracked kinematic variable estimates with known values for the synthetic sequence and with values obtained from manual digitising for the video sequences.

METHOD

The position and angle time histories (11 sets) for a half twisting forward somersault with shoulder and hip configuration changes were generated using a simulation model of aerial movement (Yeadon *et al.* 1990). A total of 10 variables were altered in this movement: 3 pelvis positional coordinates, 3 pelvis orientation angles (representing whole-body orientation), 2 arm abduction angles and 2 thigh elevation angles. Image sequences of this movement were generated (Open Inventor, SGI) from three views (front, side and top) using a NURBS (Non-Uniform Rational B-Spline) surface-based graphics representation of an 18 segment human body model (Figure 1a). The camera and lighting parameters used in the creation of the sequences were known.

An ‘Ellipsoid’ volume-based version of the 18 segment human body model was constructed to be the tracking model (Figure 1b). A total of 33 ellipsoid solids were used with each body segment comprising between one and four solids. Orientation and position were each described by three variables while body configuration was specified using 51 joint angles.

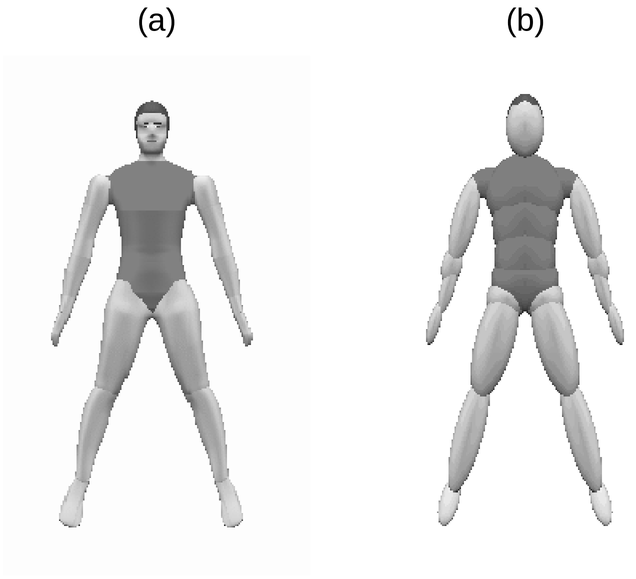


Figure 1. (a) NURBS model; (b) Ellipsoid model.

The tracking procedure, based around a ‘generate-and-test’ approach, has been described previously (Trewartha *et al.* 2003) but has the following basic structure. For each set of synchronised target images, associated model images containing the Ellipsoid human body model within simulated environments are generated and matched onto the target images. Matching is evaluated by defining a “mask” comprising the pixels corresponding to the model in the model image and identifying the corresponding pixels in the video image through the mask. A score based on the RGB colour difference between the model and target images is then calculated. Model configuration values are altered in an iterative manner to minimise the RGB difference score between the model images and target images. Final model values provide estimates for the position, orientation and configuration of the human figure at each time instant of a movement sequence.

A hierarchical procedure was used to determine the tracking variables by first optimising the six variables defining the position and orientation of the torso. Subsequently segments were added outwards along the link system. At each stage of tracking those model segments directly relating to variables being optimised were successively included in the model images generated for the model-to-target comparisons. Thus in the initial tracking of the torso no limb segments were included in the model. This approach was adopted since attempting the simultaneous improvement of all variables describing body configuration may lead an automatic tracking approach into major difficulties as the search space is relatively high and so locating a global optimum is an ill-conditioned problem.

A collection of downhill-stepping optimisation algorithms (based on the RGB difference score) were implemented to control the iteration of the model configuration to the best estimate of body posture. In early matching at each time step, algorithms that iterated a number of variables to a global optimum simultaneously (multi-parameter) were used. In later matching, algorithms that iterated a single variable at a time were used together with a final quadratic refinement of the estimate corresponding to minimum score.

The 11 frame synthetic target sequences of articulated human motion were tracked using the Ellipsoid model and employing the hierarchical structure for optimisation of all 10 motion variables. In each set of time-matched frames the pelvis variables were tracked initially followed by arm variables and then followed by leg variables. This process was repeated with successively smaller increment steps being used by the algorithms to refine the optimisation of the variables. The starting pose for each frame was defined by the extrapolated values from the optimised poses obtained for the previous two frames.

The accuracy of tracking the synthetic half-twisting somersault was assessed by comparing the values of the six tracked variables with the known values used to generate the sequences.

Two gymnasts, who gave informed consent, performed a number of aerial gymnastic movements from a trampette and were recorded by three genlocked Hi-8 video camera systems. The scene was illuminated solely by two 2000 Watt spotlights, one to the front/right and one to the right/rear of the movement space (Figure 2).

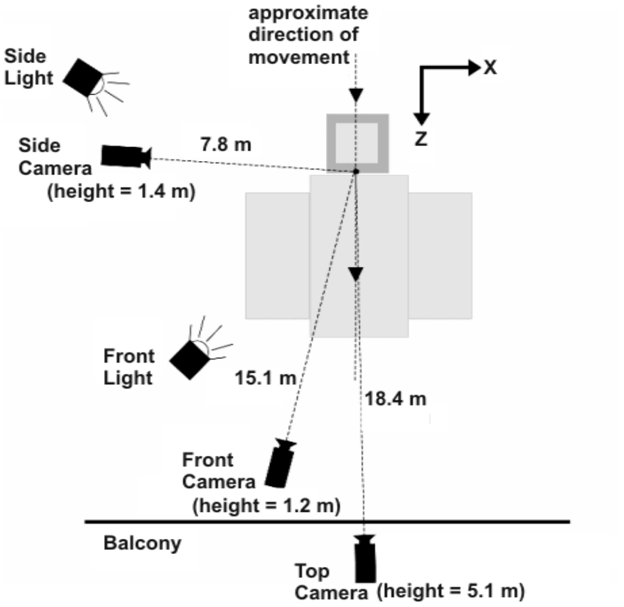

approximate direction of movement
Side Light
X
Z
Side Camera (height = 1.4 m)
7.8 m
Front Light
15.1 m
18.4 m
Front Camera (height = 1.2 m)
Balcony
Top Camera (height = 5.1 m)

Figure 2. Layout of video data collection area for aerial movements.

Parameters to be used for specifying the model cameras in the simulated environments were obtained by calibrating the video cameras via the Direct Linear Transformation (DLT) procedure, and then back-calculating the initial geometric camera parameter estimates from the DLT parameters. A constrained multi-parameter optimisation procedure (Simulated Annealing) was then used to obtain seven restricted parameters (location, orientation, focal length) available to specify the model cameras. Model lighting parameters were obtained by estimating the position and orientation of the spotlights and substituting these values into the simulated environments.

Selected aerial movements were captured via an image capture board into SGI movie format. Post-processing (separation of fields and vertical interpolation) of captured video sequences resulted in three synchronised image sequences for each movement displayed at 768 x 576 pixels at 50 images per second. The following movements were tracked: a starjump movement by the female subject (37 images), a piked forward somersault by the female subject (43 images) and a half twisting forward somersault by the male subject (45 images).

The Ellipsoid graphics model was customised to each of the gymnasts using anthropometric data to scale body segments and RGB sampling from the video images to select segment colouring (Figure 3).

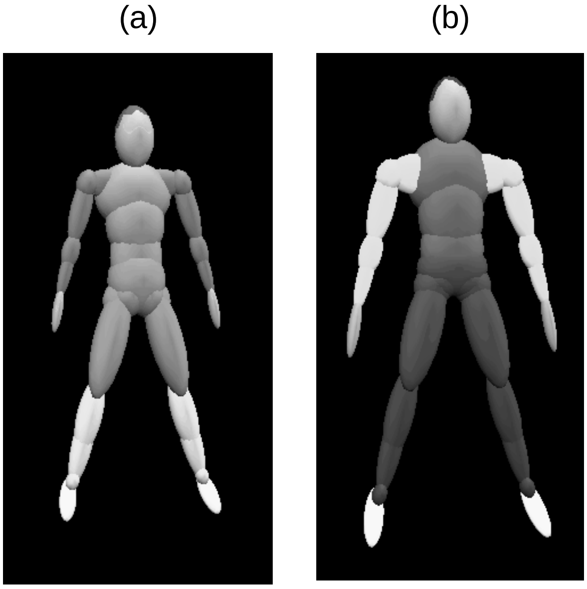


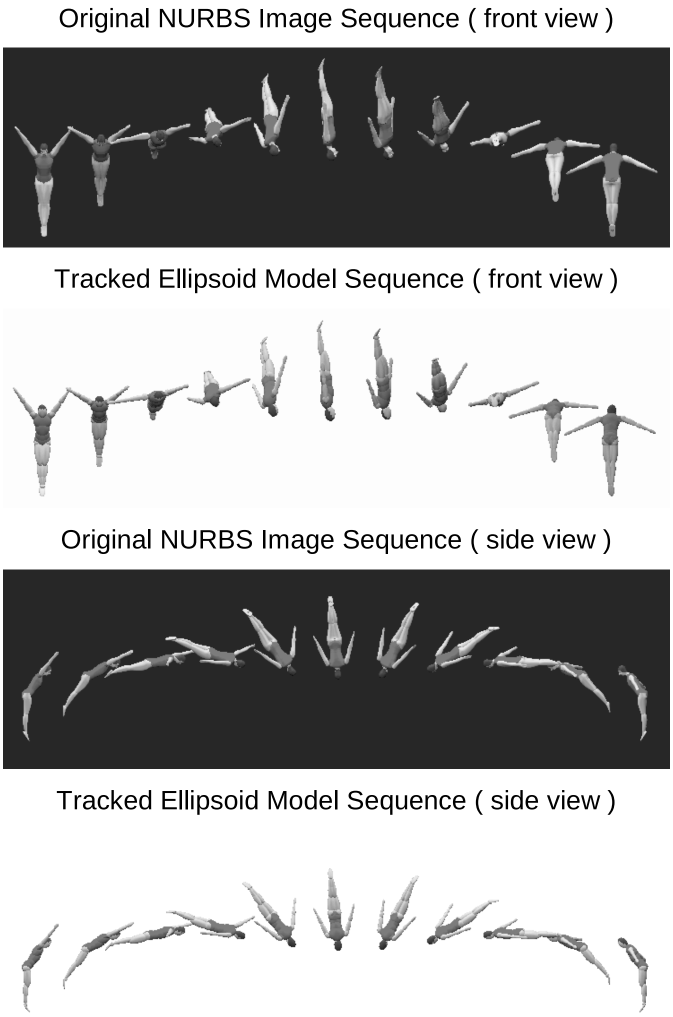
Figure 3. Model images of the Ellipsoid human body model customised for anthropometry and colour for (a) female subject and (b) male subject.

The hierarchical tracking approach was applied to the video sequences with RGB difference scores based on an equal weighting between colour differences obtained using original RGB pixel intensities and differences obtained using pixel intensities normalised for overall intensity (normalised RGB). A total of 15, 20 and 14 model variables were altered for the starjump, piked somersault and twisting somersault movements respectively. Initial model configurations were obtained from manual digitising estimates of the first video fields. Following initialisation in the first field the tracking process was fully automatic: there was no ‘boot-strapping’ procedure available to re-initialise the model configuration at any future stage. The final values from the first field were used as the initial estimates for the second field and linear extrapolation was used to estimate the initial variables in subsequent fields.

The tracking performance of the three video recorded movements was assessed by visual comparison of the original video sequences with the generated model images and by comparing estimates of the tracked variables with values obtained via manual digitisation.

RESULTS

The final model configurations of the synthetic half twisting somersault returned by the tracking procedure were used to produce graphics sequences that were compared with the target sequences from the front and side views (Figure 4).



Original NURBS Image Sequence (front view)

Tracked Ellipsoid Model Sequence (front view)

Original NURBS Image Sequence (side view)

Tracked Ellipsoid Model Sequence (side view)

Figure 4. Comparison of target and tracked image sequences from front and side views for the synthetic half twisting forward somersault movement.

The accuracy of the tracking of the synthetic movement was good with root mean square errors less than 5 mm and 1.2^o (Table 1).

Table 1. Tracking error estimates for the synthetic half twisting somersault

Variable	Range	Max Error	RMS Error
pelvis vertical	790 mm	8 mm	4 mm
pelvis A-P	980 mm	9 mm	5 mm
pelvis lateral	10 mm	5 mm	2 mm
pelvis somersault	282°	0.7°	0.4°
pelvis tilt	6°	1.4°	0.5°
pelvis twist	175°	1.0°	0.4°
L-arm abduction	130°	1.1°	0.7°
R-arm abduction	130°	0.9°	0.4°
L-thigh elevation	40°	2.9°	1.1°
R-thigh elevation	40°	2.5°	1.0°

Note: A-P = anterior-posterior, L = left, R = right.

The original video target images of the movement were compared with the generated model images and the target-through-mask images which showed the region of the target images covered by the model boundaries (Figures 5 – 7). The rms differences in kinematic estimates obtained from the automatic tracking procedure and manual digitising were calculated (Table 2). Repeated digitisation provided an independent estimate of the precision of the manual digitising process. The rms difference values obtained from repeated manual digitising trials were between 6-16 mm for position estimates, 1-3° for pelvis orientation angles, 3-5° for arm angles and 2° for leg angles, depending on the movement.

Table 2. RMS differences between estimates obtained from tracking and manual digitising

Variable	RMS Difference		
	star jump	piked som	twisting som
pelvis vertical	11 mm	34 mm	41 mm
pelvis A-P	12 mm	30 mm	39 mm
pelvis lateral	10 mm	22 mm	26 mm
pelvis somersault	1.5^o	6.1^o	6.3^o
pelvis tilt	2.9^o	6.8^o	4.4^o
pelvis twist	-	8.6^o	13.0^o
L-thigh elevation	7.4^o	6.4^o	10.2^o
L-thigh abduction	6.4^o	11.1^o	9.4^o
R-thigh elevation	1.8^o	5.9^o	8.1^o
R-thigh abduction	4.4^o	9.1^o	7.6^o
L-upperarm angle *	12.1^o	14.6^o	28.1^o
R-upperarm angle*	7.7^o	20.8^o	23.1^o
L-forearm elevation		11.8^o	
R-forearm elevation		8.5^o	
L-calf elevation		7.5^o	
R-calf elevation		5.9^o	

Note: * upperarm angle gives the difference in arm orientation after elevation and abduction.

A-P = anterior-posterior, L = left, R = right.

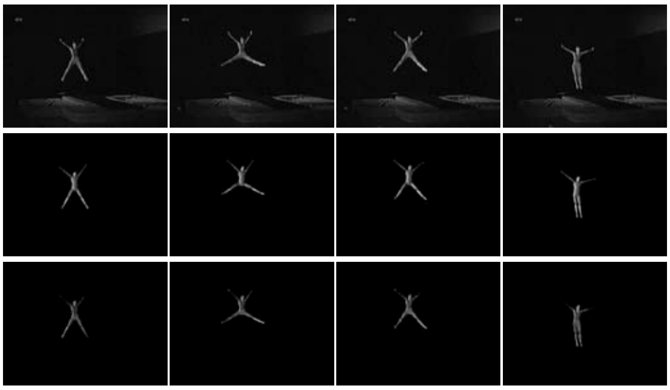


Figure 5. Video (upper sequence), model (middle sequence) and target-through-mask (lower sequence) images from the front camera view (selected fields) for the starjump movement.

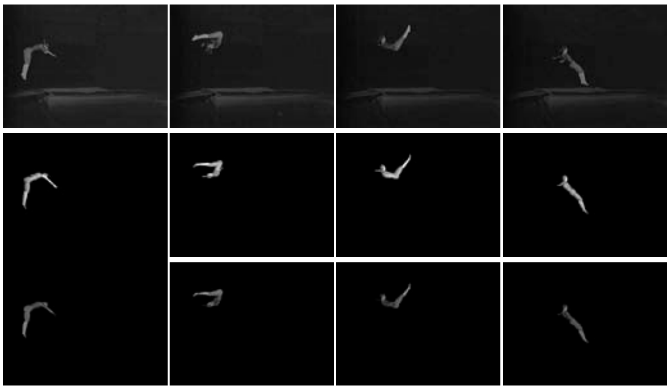


Figure 6. Video (upper sequence), model (middle sequence) and target-through-mask (lower sequence) images from the side camera view (selected fields) for the piked somersault movement.

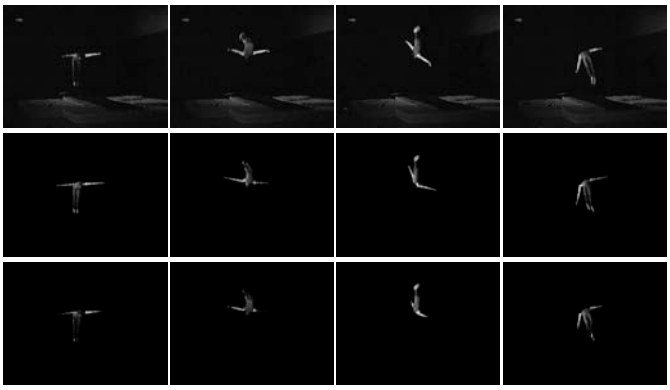


Figure 7. Video (upper sequence), model (middle sequence) and target-through-mask (lower sequence) images from the front camera view (selected fields) for the half twisting somersault movement.

There was a tendency for the time histories of the tracked variables on the right side of the body to be in better correspondence with the digitised estimates than the variables on the left side of the body (Figures 8-10). This was due primarily to the placement of video cameras and light sources within the data collection environment (Figure 2) which resulted in poor illumination of the left side of the body.

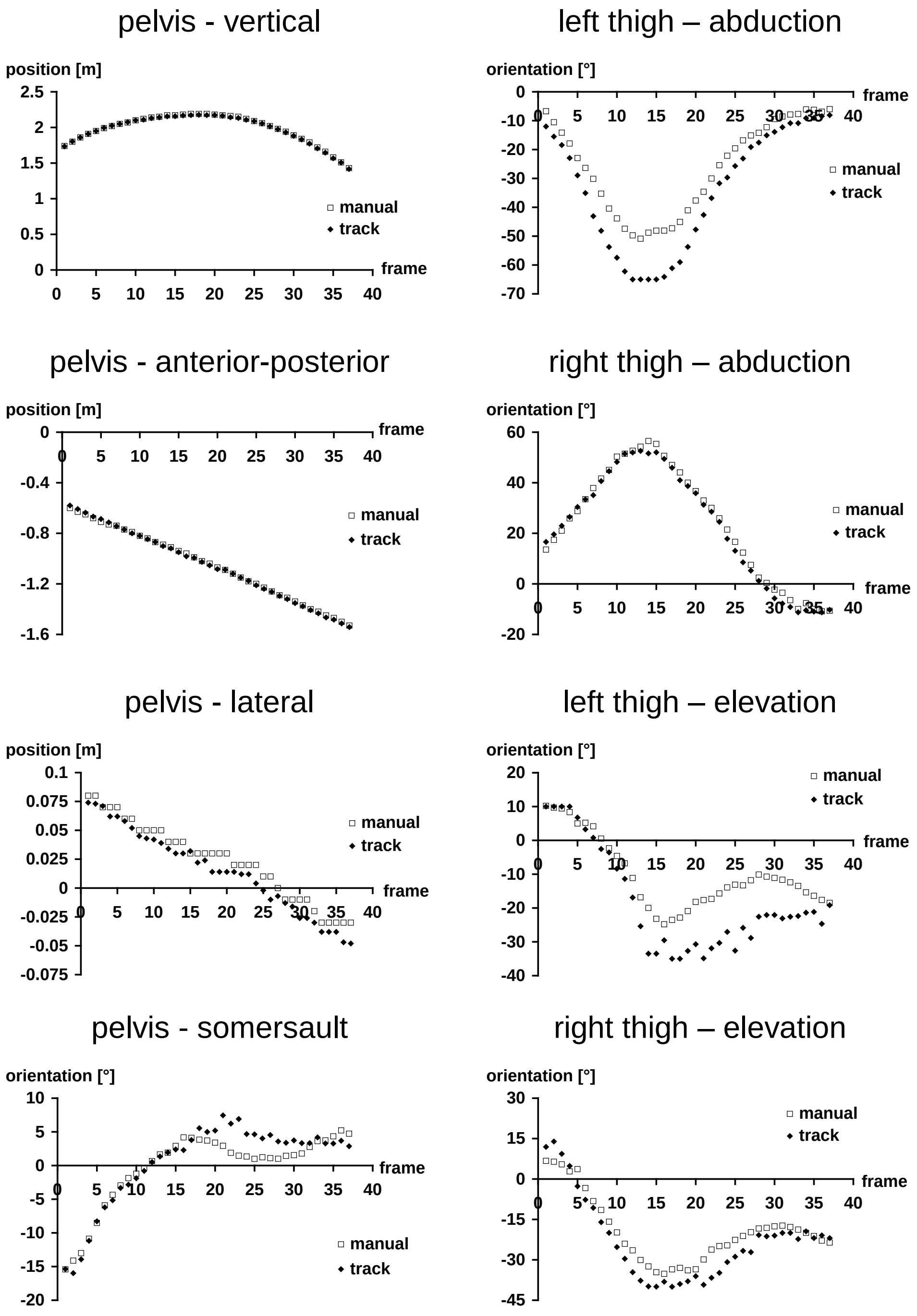
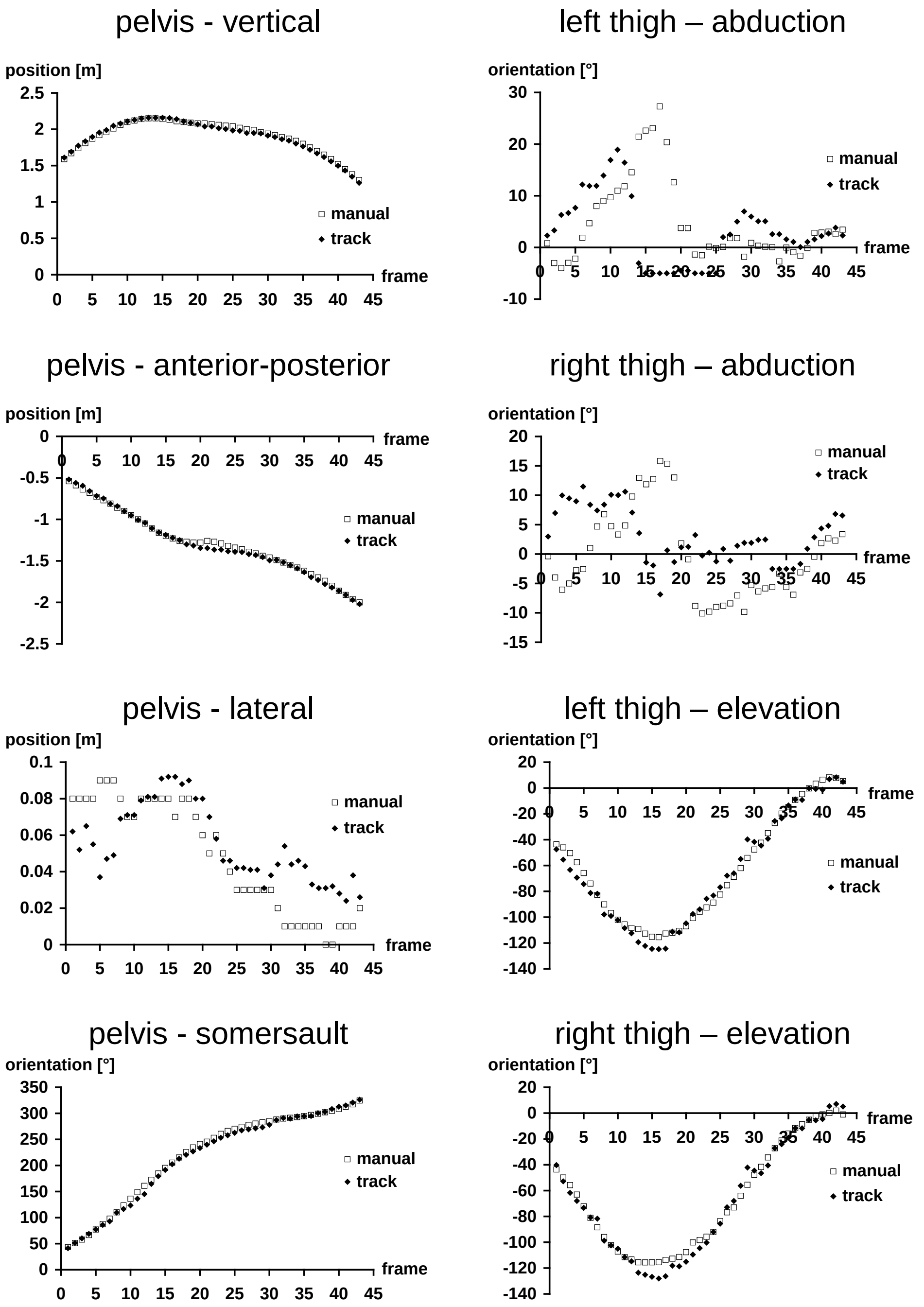


Figure 8. Time histories of the tracked variables during the starjump movement obtained from automatic tracking and manual digitising.


pelvis - vertical

left thigh – abduction

pelvis - anterior-posterior

right thigh – abduction

pelvis - lateral

left thigh – elevation

pelvis - somersault

right thigh – elevation

Figure 9. Time histories of the tracked variables for the piked somersault movement obtained from automatic tracking and manual digitising.

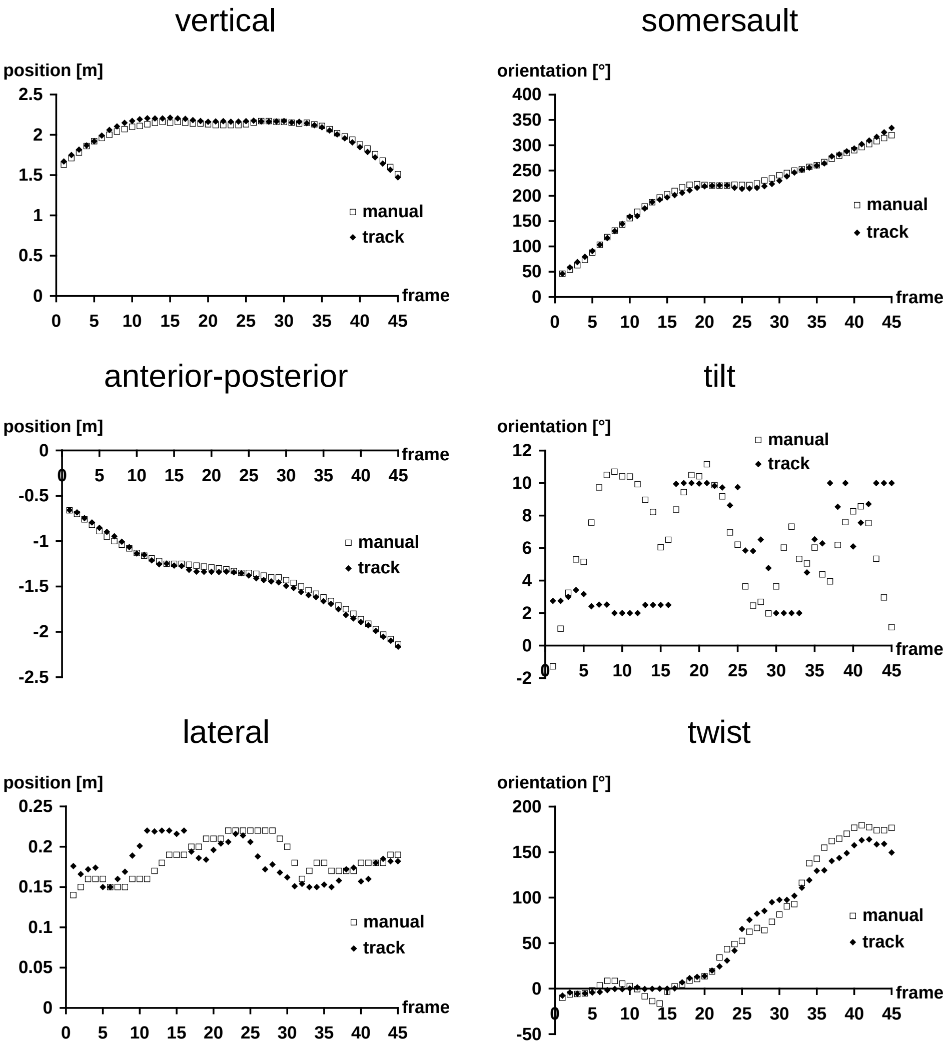


Figure 10. Time histories of the tracked pelvis variables for the half twisting somersault movement obtained from automatic tracking and manual digitising.

Although not presented, the time histories for the leg abduction and elevation variables for the twisting somersault movement had the same degree of correspondence (automatic tracking vs. manual digitising) as the leg elevation variables from the piked somersault movement (Figure 9).

DISCUSSION

The present results on synthetic image data (Figure 4 and Table 1) confirm that the model-based tracking approach is capable of returning accurate estimates when extended to the tracking of articulated movement. RMS error estimates were less than 5 mm for pelvis position estimates, less than 0.5° for pelvis orientation estimates and less than 1.2° for limb configuration angles. Despite the experiments being run on synthetically generated image data these low error estimates are impressive considering the inter-frame steps taken by some of the variables. The entire movement was covered in only 11 frames, resulting in arm angle changes of up to 70° between consecutive frames. The iterative tracking algorithms were able to cope with such large changes despite using initial estimates that were constrained to lie within 8^{o} of the final values from the previous frame. This result gives some confidence that the approach has the potential to produce very accurate estimates under good tracking conditions.

Comparing the target images to final model images for each video-recorded movement (Figs. 5-7) it is evident that the procedure was successful in tracking the prominent features of each movement. The model images obtained for the starjump and piked somersault movements exhibit extremely good correspondence with the video images while the model images for the twisting somersault movement are generally good but show periods of inconsistent tracking. Similarly, the target-through-mask images demonstrate that the registration of the model onto the human subject at each stage is also occurring with good consistency for the vast majority of time steps for each movement.

When the tracked estimates and the estimates obtained from manual digitising are compared (Table 2 and Figures 8-10) it is evident that the starjump movement seems to have been tracked with good accuracy, the rms difference values approaching the limits of the manual digitising precision. As the movement complexity increases, the rms differences also increase. The rms differences for the piked somersault movement are approximately twice as large as those observed for the starjump movement and the rms difference values for the twisting somersault movement are larger still. Generally limb angles have larger rms values than pelvis variables since the errors in estimating pelvis variables are propagated to more distal variables. On some occasions limb variables must deviate from the true values in order to maintain a good model-to-target match due to poorly estimated pelvis variables.

The results for tracking movement from video sequences are particularly encouraging bearing in mind that the data collection environment was far from ideal. The distribution of the cameras and the concentration of the light sources on one side of the body led to difficulties in estimating values for body parts on the ‘far’ side of the scene due to occlusion problems and shadowing on some regions of the body (see Figures 8 – 10). Although it was initially considered to be advantageous to brightly illuminate the subject and leave the background with little illumination, the shadowing problem for certain body parts negated any possible advantage. Moreover, since more successful tracking results have been obtained using some form of normalised colour representation rather than the intensity-based original RGB signal it is anticipated that a well-illuminated environment throughout the entire activity volume will prove more successful. Such an environment is arguably more likely to exist in a ‘real life’ setting, such as a gymnastics arena.

All vision-based tracking systems introduce a number of assumptions to make the problem tractable. The proposed tracking system is no different to others in that tracking will only be successful providing certain criteria are met. For example, the model-to-target comparison is based on a colour difference score and so it is vital that there is some colour contrast between the subject and the background and preferably between different body segments. The pose estimate is improved if body segments are not occluded from a number of camera views, if the model is a better approximation of the human form, and if reliable information about the environment (e.g. lighting and camera conditions) is available.

The ‘generate-and-test’ optimisation procedure used in this approach tends to be more computationally demanding than other pose estimation routines (Delamarre & Faugeras 2001). This procedure has been used successfully in a number of systems, however, and has been demonstrated to be a successful method in motion estimation. Due primarily to the fact that the system reported here is aimed at obtaining kinematic estimates for biomechanical analyses, robustness and accuracy considerations will initially take priority over speed of data availability. The run time for tracking the video sequences with a 600 MHz Pentium II CPU ranged from eight hours to nine hours. Processing speed may be expected to improve by a factor of around 20 when using current hardware with parallel processing of the data from each camera.

Lerasle *et al.* (1999) state that the errors involved in modelling the human form using approximate representations have a negative effect on the accuracy of pose estimation and they recommend the use of more precise and deformable representations. In contrast, Gavrila (1999) states that models need not be highly representative providing robust model matching is possible. At the present time, the 18 segment (33 solids) Ellipsoid model is considered sufficiently complex for tracking human movement from video images. Model segments have been sized according to anthropometric measurements and coloured according to a sampling of the video images. Future elaboration of the method may be necessary if the approach is to be used for tracking individuals where no information (other than video) is available on their physical characteristics or if it is felt necessary to update the RGB content of model segments based on the most recent image data during tracking.

At present the approach has run into some inconsistent matching when tracking more complex movements, particularly twisting motions. Longitudinal rotations are notoriously difficult to estimate due to the fact that little image change results from relatively larger changes in twist angle. One step that may be expected to result in immediate improvements in this regard would be to increase the number of cameras from the three used in the present study to provide more redundancy of data. Additionally patterned clothing may be used to improve the ability of the system to track torso variables accurately.

This study has presented successful tracking results on a number of aerial movements including whole-body somersaulting and twisting rotations which have not been tracked in any other studies. The tracked estimates have been evaluated against kinematic estimates obtained using an alternative method, a step that is often overlooked or avoided in tracking studies. The present results demonstrate the complexity of tracking possible using a relatively straightforward approach to the problem with only a single image cue and minimal image processing. This provides a sound basis for future development. Further additions are required before this tracking system will provide a genuine solution for biomechanics research. It is envisaged, however, that this model-based multi-view colour matching approach may form the basis of such a system.

REFERENCES

Bregler, C. & Malik, J. (1998) Tracking people with twist and exponential maps. IN: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8-15. IEEE Computer Society, Los Alamitos, CA.

Chen, Z. & Lee, H.J. (1992) Knowledge-guided visual perception of 3D human gait from a single image sequence. *IEEE Transactions on Systems, Man and Cybernetics*, **22**, 336-342.

Delamarre, Q. & Faugeras, O. (2001) 3D articulated models and multiview tracking with physical forces. *Computer Vision and Image Understanding*, **81**, 328-357.

Gavrila, D.M. (1999) The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, **73**, 82-98.

Gavrila, D.M. & Davis, L.S. (1996) 3-D model-based tracking of humans in action: A multi-view approach. IN: *Proceedings of the IEEE Computer Society Conference on Computer Visions and Pattern Recognition*, pp. 73-80. IEEE Computer Society, Los Alamitos, CA.

Lerasle, F., Rives, G. & Dhome, M. (1999) Tracking of human limbs by multiocular vision. *Computer Vision and Image Understanding*, **75**, 229-246.

Rohr, K. (1994) Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, **59**, 94-115.

Trewartha, G., Yeadon, M.R. & Knight, J.P. (2003) Colour based rigid body tracking using three-dimensional graphics models. *Sports Engineering* (in press).

Yacoob, Y. & Davis, L.S. (2000) Learned models for estimation of rigid and articulated human motion from stationary or moving camera. *International Journal of Computer Vision*, **12**, 5-30.

Yeadon, M.R., Atha, J. & Hales, F.D. (1990). The simulation of aerial movement - IV. A computer simulation model. *Journal of Biomechanics*, **23**, 85-89.