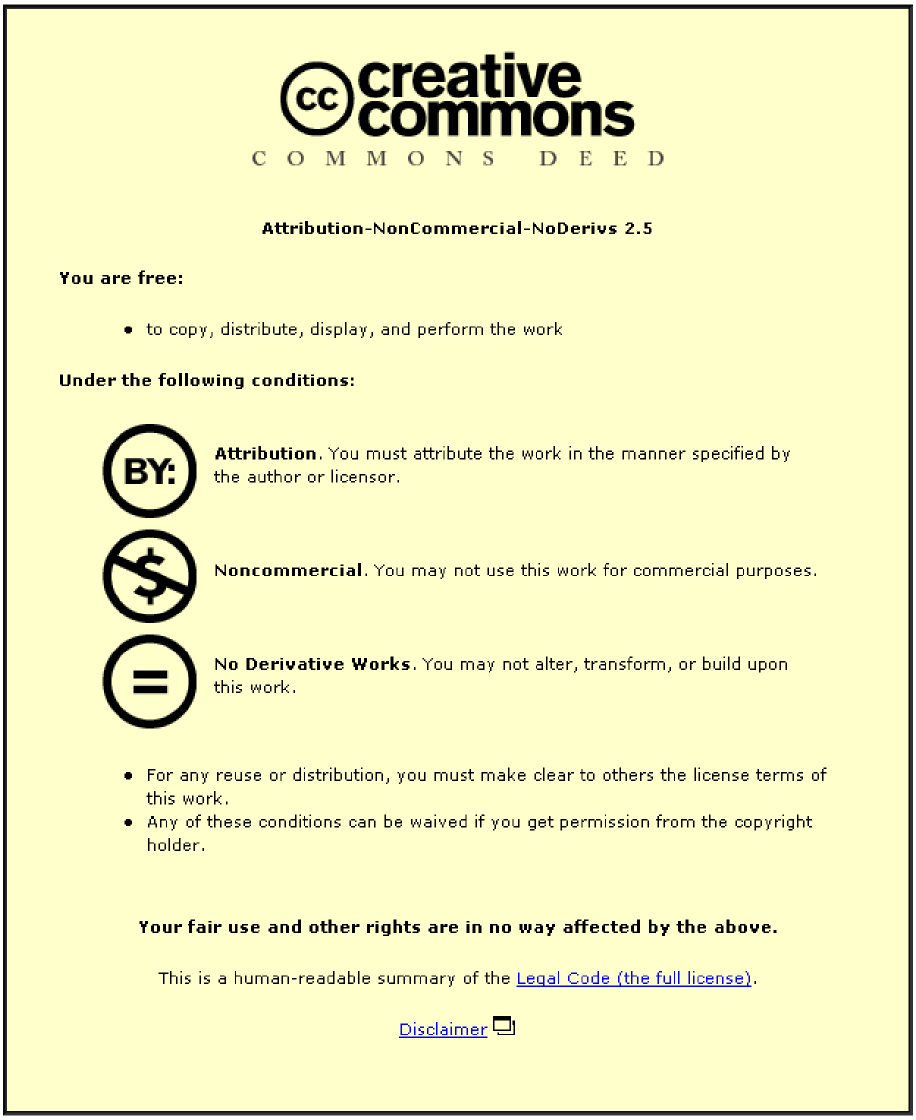


Sports Engineering 6, 139-148, 2003

TECHNICAL NOTE

Colour based rigid body tracking using three-dimensional graphics models

Grant Trewartha, Maurice R. Yeadon and Jon P. Knight

G. Trewartha Department of Sport and Exercise Science, University of Bath, UK
M.R. Yeadon and J.P. Knight School of Sport and Exercise Sciences, Loughborough University, UK

ABSTRACT

This paper introduces the first stage of a new model-based approach to three-dimensional (3D) human movement tracking. A 'generate-and-test' matching procedure was adopted by matching rendered images of a 3D computer graphics model of the human body to target images of rigid body motion. The set of pixels to be compared were just those corresponding to the model of the body in the rendered images. The matching criterion to optimise model position and orientation was based on the minimisation of the RGB colour difference between generated model images and associated target images. The method was able to track synthetic image sequences of a half twisting somersault accurately with root mean square errors of less than 5 mm and 0.3° for position and orientation estimates respectively. The suitability of the proposed approach for rigid body motion tracking was supported by additional tracking experiments on video image sequences of 'wooden cross' trajectories. Comparisons of tracked estimates against manual digitising estimates returned relatively small rms difference values on both side somersault and twisting somersault movements. The proposed approach has the potential to track video images of a human torso using a rigid body model and hence to track articulated movements by successively adding segments to the model in a hierarchical manner.

keywords; video, image, tracking, colour, rigid body, three-dimensional

INTRODUCTION

Three-dimensional (3D) rigid body motion recovery is often accomplished using multiple camera automated marker tracking systems. Such automatic motion analysis systems are capable of providing accurate 3D motion data but have inherent limitations such as relative marker movement and non-coincidence of markers with supposed anatomical location. Manual digitising systems can operate without markers but are labour intensive. There are situations, such as Olympic competition, where markers cannot be used and in these situations an automatic marker-free system would have considerable advantages.

Numerous studies have attempted to recover human motion parameters from images without the use of markers (Bregler & Malik 1998; Plänkers & Fua 2001). However, procedures that can robustly track unconstrained 3D whole-body human movement have still to be developed. The first step towards the ultimate goal of 3D articulated human motion tracking is to develop a method that is capable of tracking movement under more constrained conditions. One means of constraining the problem is to track rigid body motion, which has fewer degrees of freedom (DoF). If a successful approach permits hierarchical extension from a single rigid body to a system of linked rigid bodies then it may be possible to track articulated human movement.

For motion estimation, specific vision-based approaches may use image data exclusively by detecting corresponding image features over consecutive frames or may employ a model-based procedure where model features are matched to image features at each stage and the model state is used as an estimate. Irrespective of the approach taken, all vision-based tracking techniques must exploit some of the information held within the image sequences.

Many tracking approaches now employ low-level features, such as image intensity, where matching is performed on a global basis. Colour information can provide an additional cue for matching so colour images have been used to aid tracking procedures (McKenna et al. 1999). The RGB (red-green-blue) colour intensities of an object image are influenced by illumination conditions and so colour representations invariant to illumination intensity may be beneficial.

Attempting 3D pose (position and orientation) estimation solely using image data is extremely complicated and often unreliable so shape models are often introduced to assist pose recovery. Model-based rigid body tracking methods may use edge features (Li & Wang 1999), point features or region features (Khotanzad & Liou 1996).

Hel-Or & Werman (1995) used a feature point model-based approach to locate a planar object in four different poses to within 0.009 m and 1.1° of translation and rotation respectively, with ranges of 0.49 m and 46.0° respectively. Li & Wang (1999) matched model image vertices and high contrast synthetic image vertices of known polygon objects to estimate pose for single images to within 0.009 m / 1°. For six DoF tracking over longer synthetic image sequences (50 frames) the method worked well on movements with small inter-frame motion and constant angle variance but tracking was lost as angle changes and accelerations increased. Employing multiple camera views to provide some redundancy in the estimation, Stephens (1990) matched model and image edges to track an object as it was moved by a robot arm with rms errors of 0.007 m and 3.4° during stable tracking. In general, results on six DoF tracking of synthetic images are quite good but reports of successful tracking using real image data are scarce.

The aim of this paper is to assess a model-based method that exploits the properties of colour images to track the motion of rigid bodies from the perspective of developing an automatic approach to human motion tracking.

METHOD

A human body model was developed using Open Inventor software running on an SGi platform. The 18-segment link model comprised: feet, shanks, thighs, pelvis, torso, chest, shoulders, upper arms, forearms, hands and head. Body position was specified by the 3D coordinates of a point at the top of the pelvis segment while orientation was defined by angles of somersault, tilt and twist, comprising successive rotations of the pelvis about a lateral horizontal axis, a frontal axis and a longitudinal axis respectively. Each of the 17 joints in the model had three degrees of freedom corresponding to three configuration angles describing the relative orientation of adjacent segments. In all, the pelvis position coordinates and 54 angles completely describe the pose of the human body model. If the model maintains a fixed body configuration for a complete movement then body pose can be described solely in terms of the 3D position of the pelvis and the 3D orientation angles of the pelvis.

A 3D computer graphics human body model for the target images was created using a NURBS (Non-Uniform Rational B-Spline) surface-based representation (Fig. 1a). A whole body surface scan was obtained on one individual using the LASS system to provide the surface coordinates for the graphics model (Jones & West 1989). A movement sequence was generated containing 11 poses of the NURBS human body model executing a half twisting forward somersault. The time histories of pelvis positions and orientation angles for this movement were obtained using a computer simulation model of aerial movement (Yeadon et al. 1990). Synthetic image sequences were generated from three views: front, side and top with known camera and lighting parameters used in the production of the target image sequences. The human graphics model maintained a fixed configuration throughout the movement and thus behaved as a rigid body.

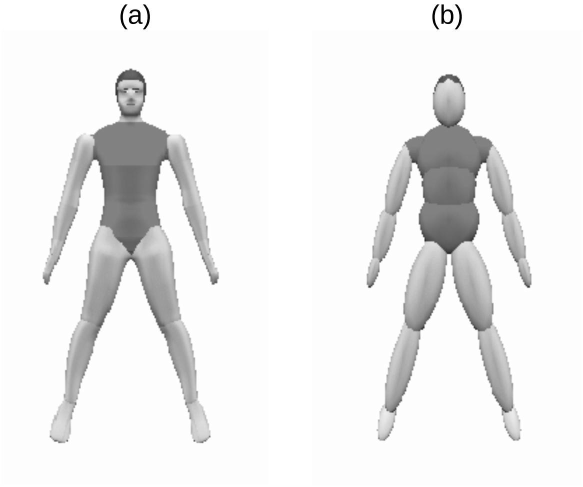


Figure 1. (a) NURBS model, (b) Ellipsoid model.

In order to track the image sequences a more simplistic volumetric 3D human body model (Fig. 1b) was constructed using a single ellipsoid to represent each of the 18 body segments in such a way that the ‘Ellipsoid’ model had the same joint centre locations as the ‘NURBS’ model and similar segment dimensions. The colouring of the Ellipsoid model segments was identical to that of the NURBS model used in the creation of the target image sequences. However, due to the different shape representations, the positions at which colours changed were slightly different and the surface geometry produced differences in RGB pixel intensities for corresponding body points. Since the target image sequences were computer-generated it was possible to create model images with identical environment information (camera and lighting parameters).

The estimation procedure was based around a ‘generate-and-test’ approach, where the aim was to establish the model pose whose images were most similar to the target images in terms of the rms RGB (red-green-blue) pixel colour difference. A “mask” comprising the pixel locations for which the RGB value was non-black in the model images was used to select the portion of the target image for the difference calculation (Fig. 2).

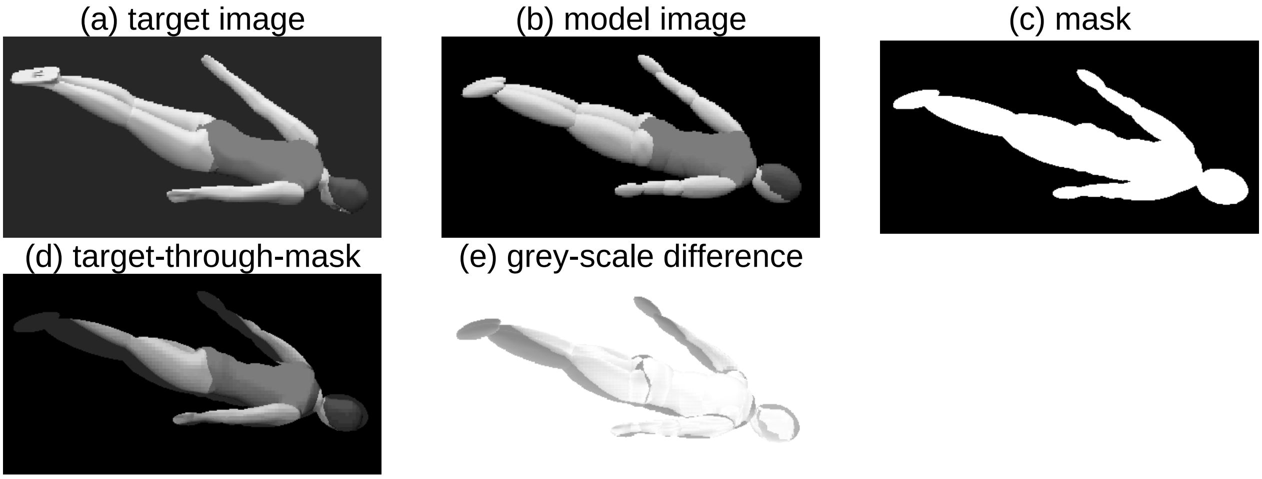


Figure 2. Images representing the model-to-target matching process:- (a) target image [synthetic or video], (b) image generated of the model in a particular pose from a simulated environment, (c) mask produced from all non-black pixels in the model image, (d) region of the target image visible through the mask, (e) evaluation of the match with increasing grey-scale representing good correspondence between target image pixels and model image pixels.

The 11 frame target sequence (three camera views) containing the NURBS model executing one somersault with half a twist and considerable vertical (0.79 m) and anterior-posterior displacement (1.0 m) was tracked using a simple downhill stepping algorithm to reduce the RGB difference score between model and target images and to update the pose (6 variables) of the Ellipsoid tracking model. In order to minimise the RGB difference score each variable was optimised on an individual basis in a separate cycle of the algorithm and five 'loops' with decreasing step lengths were used to optimise the six variables for each time instant. The loops progressed with decreasing step lengths of 50 mm, 30 mm, 20 mm, 10 mm, 10 mm for pelvis position increments and 5°, 3°, 2°, 1°, 1° for pelvis orientation increments. The downhill stepping algorithm stopped after two consecutive increases in the RGB score and a V-algorithm was used to find the intersection of straight lines through the last two pairs of points allowing final model estimates to lie within the step intervals.

The RGB score was evaluated around the target values by allowing each of the six pose estimates to vary in turn. This convergence analysis allowed an assessment of how accurate the initial estimates of the variables had to be in order for the downhill stepping algorithm to be successful. The convergence analysis was conducted for Ellipsoid on Ellipsoid matching (simulating a tracking model that was a perfect representation of the target object) and for Ellipsoid on NURBS matching (simulating a tracking model that was a simplified representation of the target object).

This tracking approach was extended to rigid body tracking from colour video image sequences of an airborne wooden cross. The major extension to the method was to establish camera and lighting parameters to be used in creating the simulated environments for generating model images. A wooden cross of known dimensions was constructed from two wooden rectangular prisms, painted a bright yellow colour and was thrown in multiple trials within a volume in side somersault and twisting somersault motions. Two genlocked Hi-8 video camera recorder systems (left and right) recorded the wooden cross trajectories, viewing along horizontal axes. A third mini-DV camera recorder system, which was not genlocked, viewed the scene from above the object motion (Fig. 3). The shuttering of the third camera was manually adjusted using a timing clock to be 'in-phase' with the other two cameras to allow synchronous images to be taken from all three cameras. Two trials were selected for automatic tracking: a side somersault trial and a half twisting somersault trial.

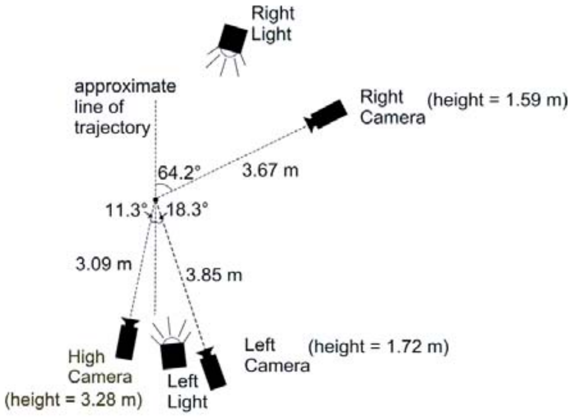


Figure 3. Plan view of video data collection of wooden cross trajectories.

Video images were captured at a resolution of 768 x 576 pixels and at 50 Hz (via separation of fields and interpolation of missing horizontal lines). The side somersault sequences contained 39 images for each camera view while the half twisting somersault sequences comprised 34 images.

To track the wooden cross sequences a ‘YellowStick’ graphics model was created using appropriately sized solids and colouring taken from the RGB content of the wooden cross in selected video images. Model camera parameters for the simulated environments were obtained via back-calculation of 11 DLT parameters to provide estimates for the camera geometric parameters (location and orientation) followed by a constrained optimisation procedure (Simulated Annealing) to specify camera set-up using the restricted seven parameter set available in the simulated environments of the Open Inventor software. The position and orientation of two 2000 Watt spotlights illuminating the scene were measured so that the lighting conditions could be replicated in the simulated environments. Since it was difficult to replicate the lighting conditions precisely the RGB values of both the video and simulated images were intensity-normalised prior to calculating RGB differences.

As for the tracking of the synthetic sequences, all six degrees of freedom of the wooden cross pose were estimated in each video image. An initial YellowStick model pose was obtained manually for each sequence by producing graphics images of the model in different 3D poses using the appropriate simulated environments until these images were a good approximation of the pose of the wooden cross in the first field of the image sequence. This approximation was used as the pose estimate to initiate tracking. Initial values to begin pose estimation in subsequent fields were obtained by linearly extrapolating model values using the estimates from the previous two fields. The image sequences were tracked by the YellowStick model using seven loops of the downhill stepping algorithm (with decreasing step lengths).

Unlike synthetic sequences, real image data does not allow the comparison of tracking estimates with known criterion values since the actual pose values cannot be acquired. Instead the pose values were compared with those obtained from manual digitisation. Additionally the target image was viewed through the mask corresponding to the model image pixels as a measure of how well the tracking located the cross image.

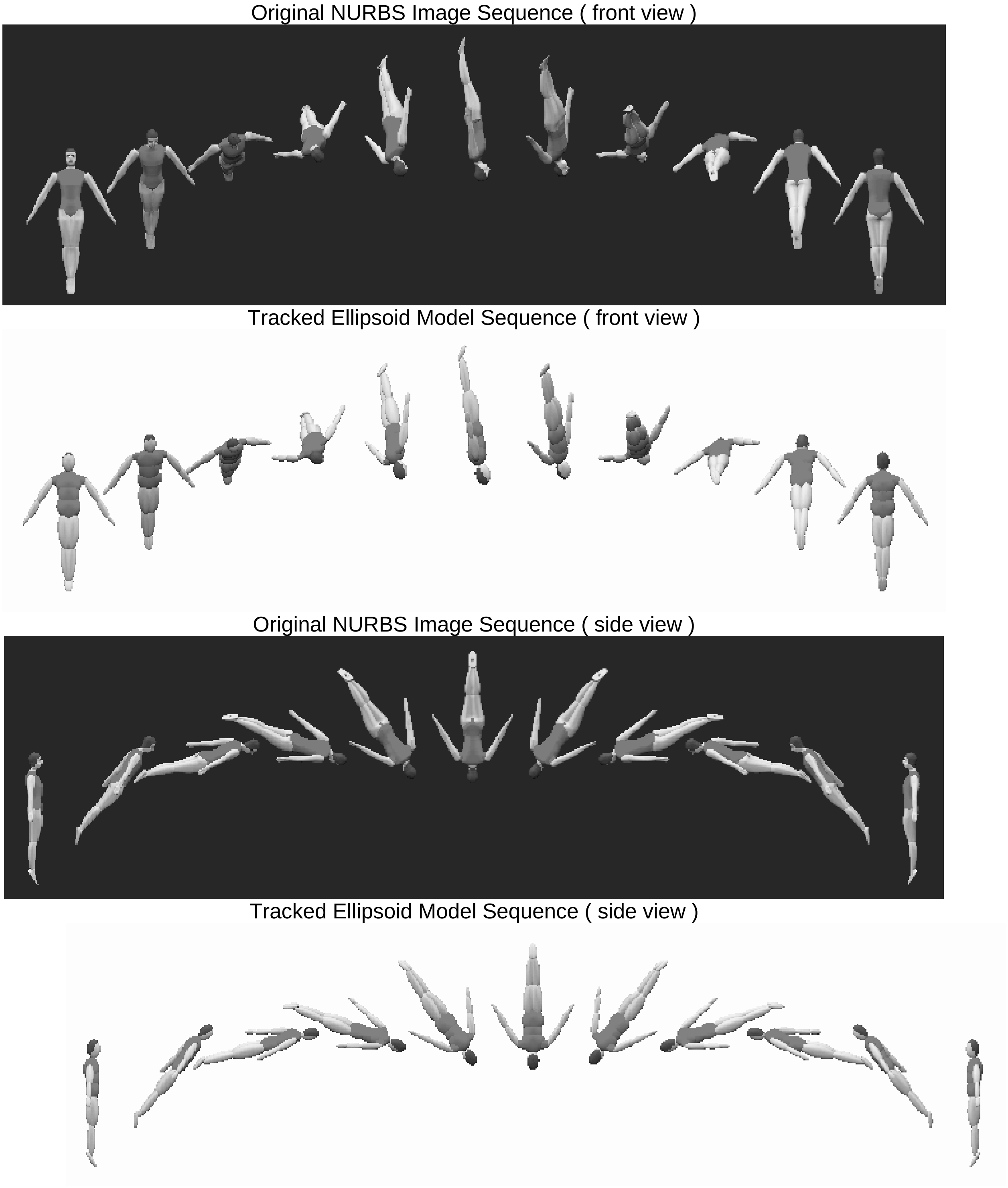
RESULTS

The accuracies of tracking estimates for the human model half twisting somersault synthetic sequence were better than 10 mm and 1^o (Table 1).

Table 1. Root Mean Square errors for tracking estimates for the synthetic half twisting somersault

Variable (pelvis pose)	Range of Motion	Maximum Error	RMS Error
vertical	790 mm	8 mm	4.8 mm
A-P	1000 mm	8 mm	4.5 mm
lateral	10 mm	3 mm	1.7 mm
somersault	360°	0.4°	0.2°
tilt	8°	0.7°	0.3°
twist	180°	0.6°	0.3°

The final model pose estimates returned by the tracking procedure were used to generate 11 frame image sequences of the movement for comparison with the original NURBS target image sequences to assess similarity (Fig. 4).



Original NURBS Image Sequence (front view)

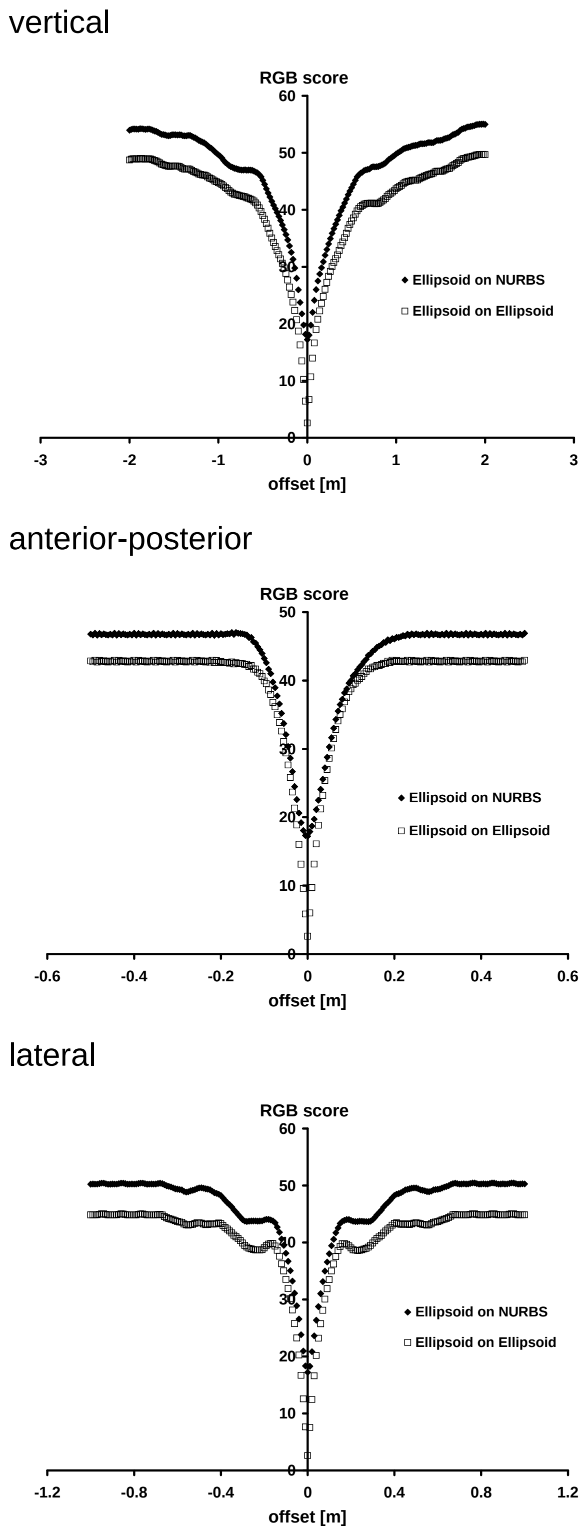
Tracked Ellipsoid Model Sequence (front view)

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 behaviour of the RGB score as a function of each pose variable was determined (Fig. 5 and 6).



vertical

anterior-posterior

lateral

Figure 5. Convergence functions for translation variables.

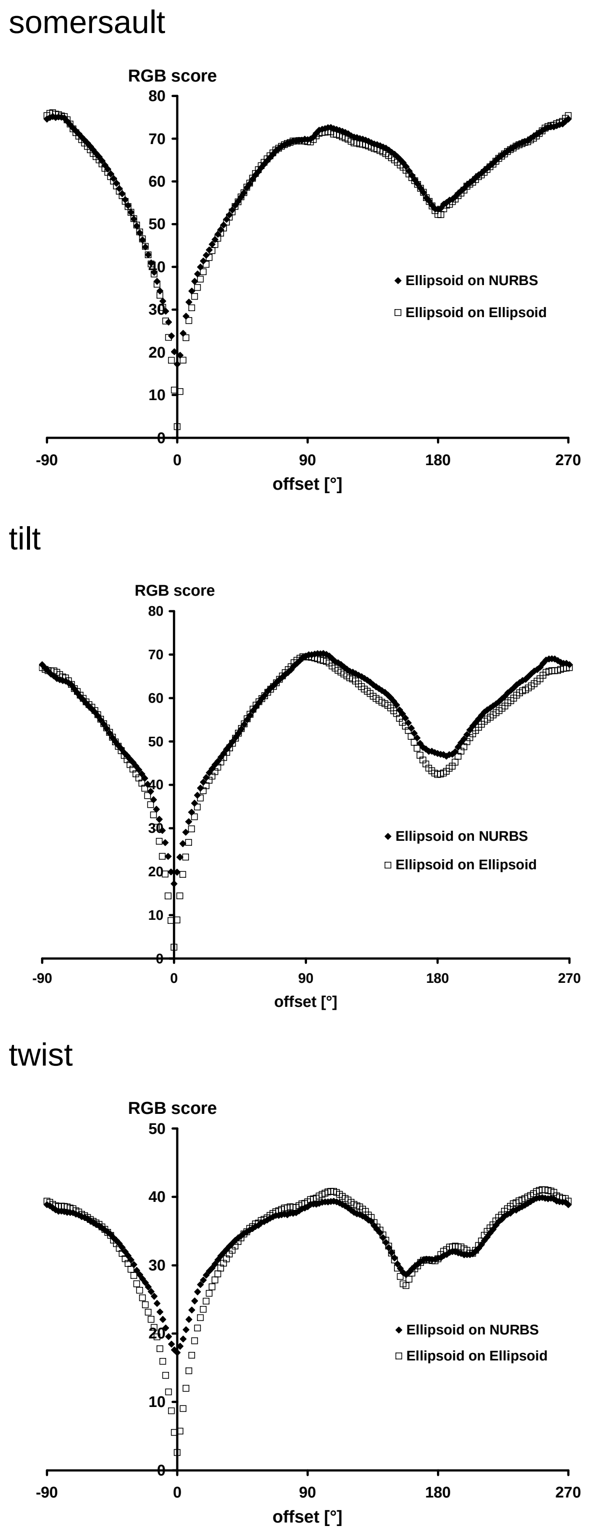


Figure 6. Convergence functions for rotation variables.

A measure of the success of the tracking of the video sequences is demonstrated by target-through-mask images for selected fields from all three camera views of the side somersault wooden cross image sequence (Fig. 7) and the half twisting somersault wooden cross sequence (Fig. 8).

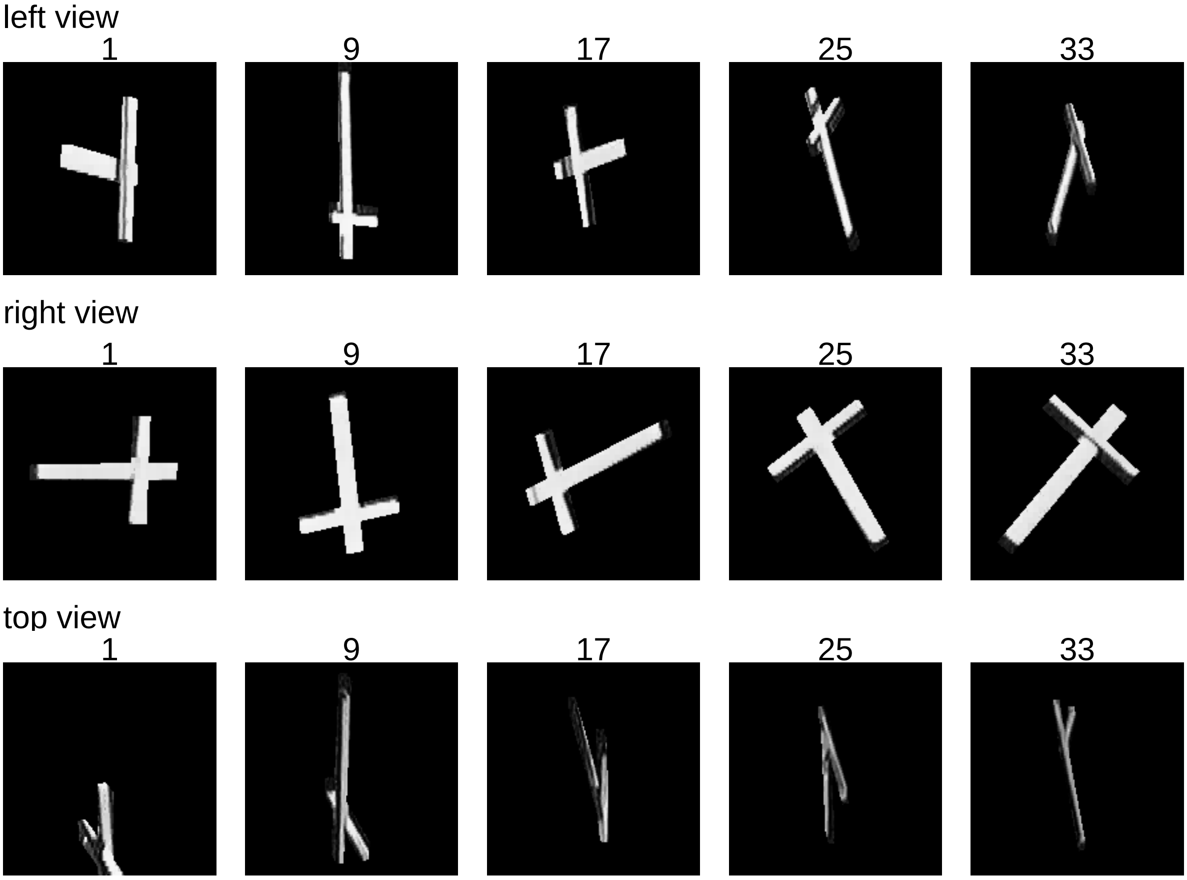


Figure 7. Target-through-mask images for selected fields of the wooden cross side somersault movement.

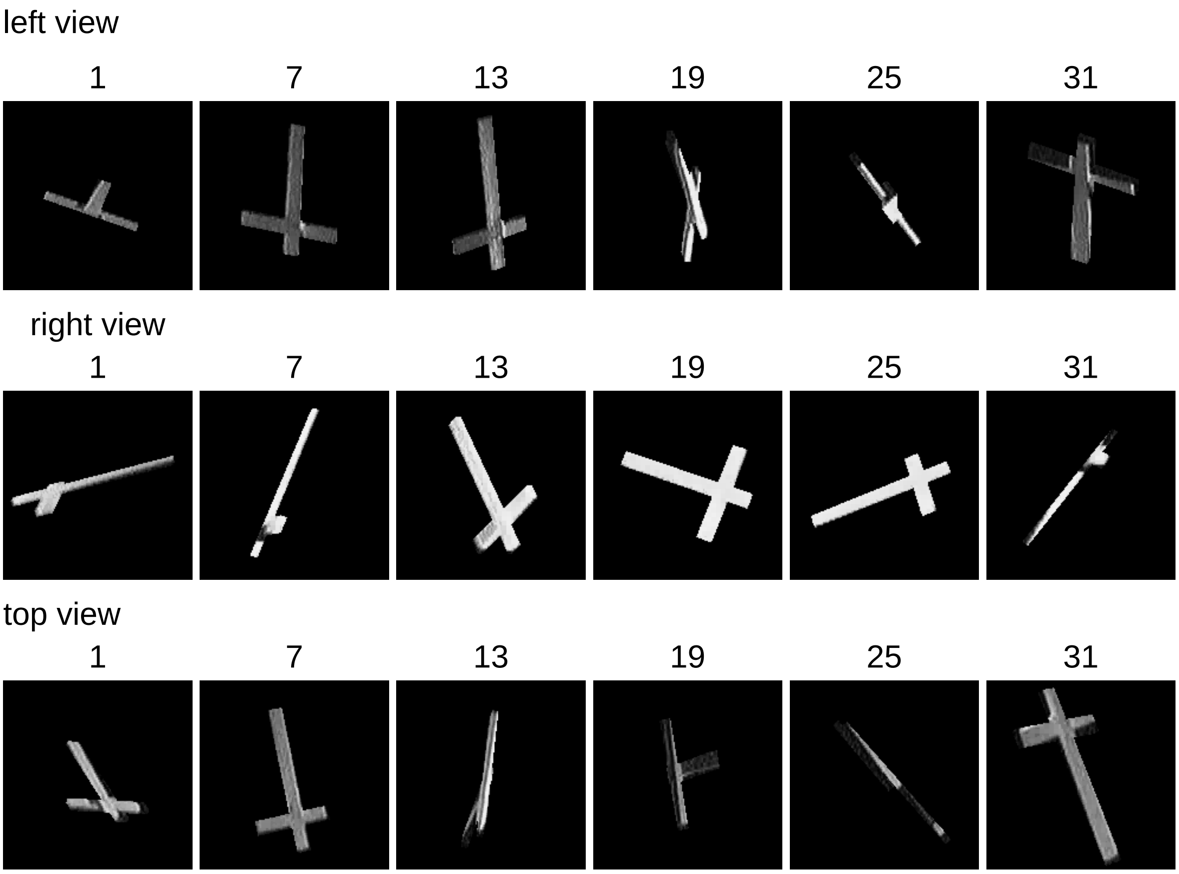
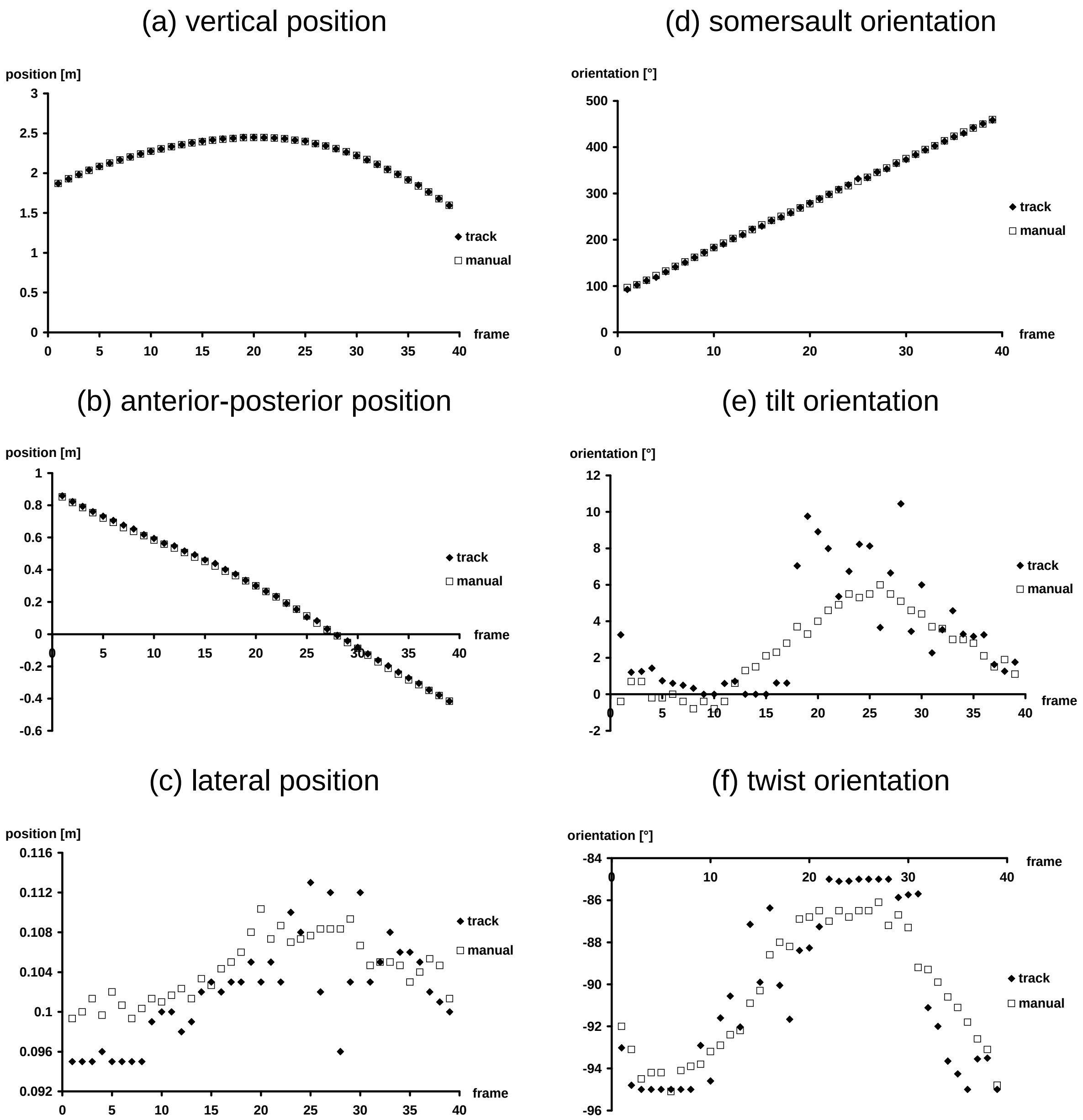


Figure 8. Target-through-mask images for selected fields of the wooden cross half twisting somersault movement.

The 3D pose estimates obtained from the automatic tracking procedure were also compared with estimates obtained by manual digitising and 3D reconstruction techniques for the side somersault sequence (Fig. 9) and the half twisting somersault sequence (Fig. 10).


(a) vertical position

(b) anterior-posterior position

(c) lateral position

(d) somersault orientation

(e) tilt orientation

(f) twist orientation

Figure 9. Comparison of pose estimates obtained from tracking and manual digitising methods for the wooden cross side somersault movement.

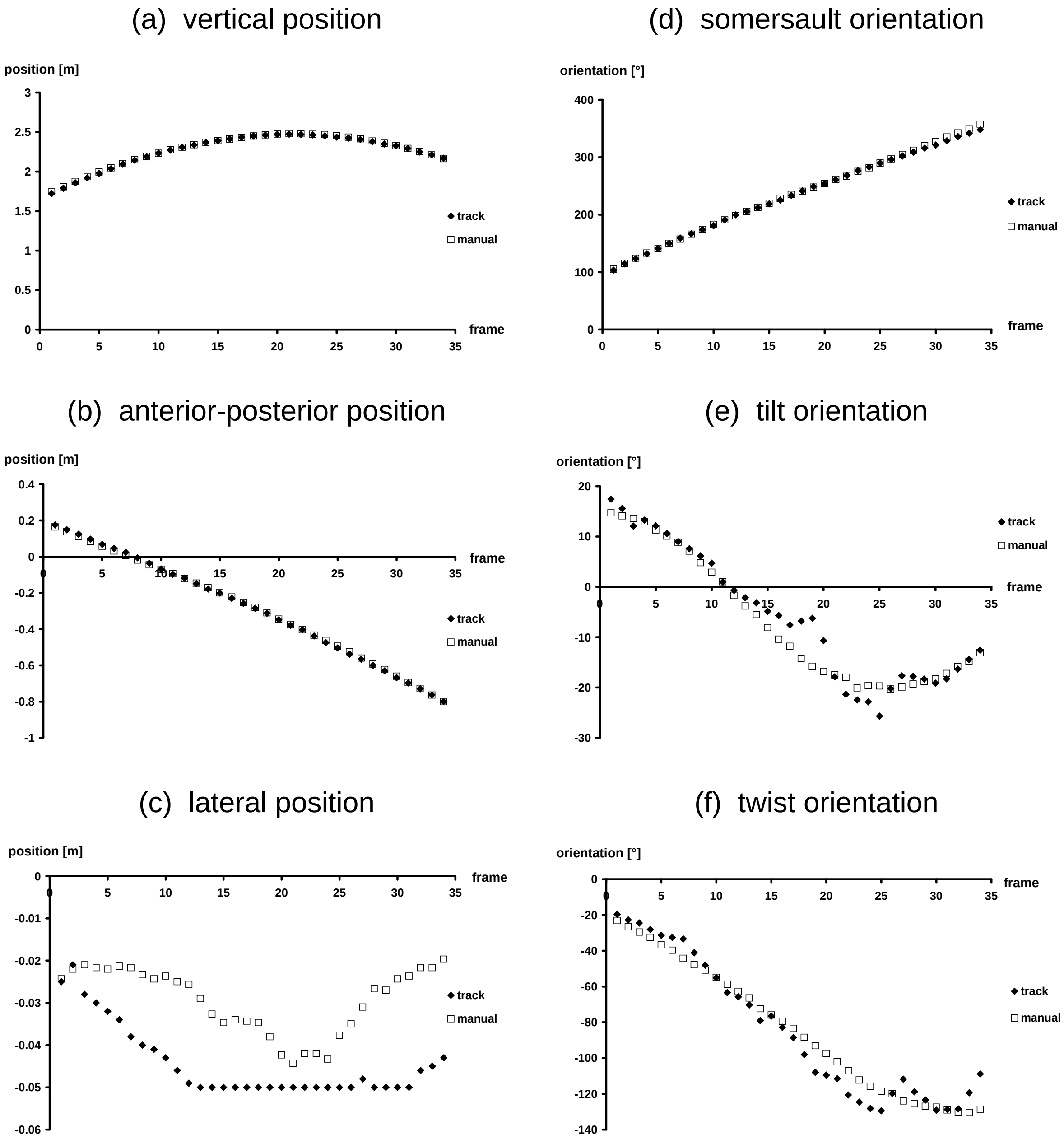


Figure 10. Comparison of pose estimates obtained from tracking and manual digitising methods for the wooden cross half twisting somersault movement.

DISCUSSION

The functions obtained from the convergence analysis (Fig. 5 and 6) led to the use of the V-algorithm to iterate model pose to the pose of the target object. Providing that step length increments in model pose are large enough to avoid local minima, it appears that the RGB matching criteria should allow iteration to a good estimate from relatively relaxed limits using the downhill stepping algorithm and the intersection of two straight lines (V-algorithm) to improve matching resolution. This finding persists even with the Ellipsoid on NURBS matching which is an encouraging sign for future matching onto video images.

The final tracking estimates returned from the analysis of the NURBS image sequence demonstrate the accuracy of the method. Over the 11 frames of the sequence the 3D position of the pelvis was located to within less than 5 mm rms error in all directions and pelvis orientation angles were estimated to within 0.3° rms error for all rotations. Maximum observed errors for any single frame were approximately twice as large as the rms errors. While this tracking was conducted on synthetic image sequences the tracking estimates nevertheless hold promise as the model used for matching was a more simplified representation of the human form than the model used in the creation of the target images. Moreover, executing a complete movement in only 11 steps means that considerable inter-frame movements existed for some of the variables, particularly somersault rotation, twist rotation and to a certain extent anterior-posterior translation. This aspect confirms the robustness of the tracking procedure to iterate to an accurate pose given only approximate initial frame estimates, since no inter-frame extrapolation was employed for the synthetic tracking.

Experiments on the wooden cross image sequences demonstrated that the proposed procedure was capable of tracking the motion of a rigid body from video sequences. The rms differences between the automatically obtained tracking estimates and estimates from an established manual digitising method were less than 10 mm and 3° in all three directions for the side somersault sequence and less than 17 mm and 9° for the half twisting somersault sequence. Residual errors, particularly for the twisting sequence could be attributed to instances throughout the movement where the pose of the model led to the generation of model images with very few object pixels. Consequently the masks for comparisons were small and limited information was available to localise pose estimates accurately.

With all vision-based tracking approaches it is important to consider the trade-off between accuracy of tracking and the number of iterations (computation cost). In this paper, the accuracy of tracking a NURBS-based target image sequence with a simple Ellipsoid model has been established. At present the method uses about 20 iterations per variable in each field to optimise the match between video and model images. It is apparent that while the Ellipsoid model is a simpler representation of the human body than the NURBS model it is nevertheless fairly realistic compared with even simpler representations such as 2D contour models. Consequently, it is speculated at this stage that the Ellipsoid representation will continue to prove successful when tracking human motion from video image sequences.

In the tracking experiments presented so far, both intensity-based (raw RGB in synthetic tracking) and intensity-normalised (normalised RGB in video tracking) colour representations have been used. It is apparent that illumination changes and out-of-plane object rotations influence raw RGB pixel values. This means that if lighting conditions are not well modelled in the simulated environments then RGB variance may lead tracking into local minima. If, however, lighting conditions are well modelled then, as suggested by Golland & Bruckstein (1997), the variance in pixel intensity can be used as an additional cue for motion estimation. It is likely that the choice of colour representation will vary depending on the recording environment and so a system that incorporates a sensible weighting between the two extremes will be beneficial.

One other aspect of the tracking approach that can potentially be altered to reduce computation cost is the implementation of alternative tracking algorithms. At present the downhill stepping algorithm has proved competent at converging towards good estimates although a drawback is the fact that it deals with each variable sequentially. This means that in the early loops of a frame, new estimates are found for variables when other model variables are still significantly deviated from their true estimates, increasing the risk of becoming stuck in local minima. This situation could be rectified by the simultaneous iteration of all variables to optimum estimates, where all variables gradually increment towards the best match together. For instance, employing a simplex optimisation algorithm on an n-dimensional surface would be suited to this purpose.

While this procedure has proved successful in tracking rigid body motions it will need to be extended to accommodate movements involving articulated motion in order to track human movement. For movements with changing body configuration the position and orientation of a segment is a function of all the configuration angles back through the link system. In order to avoid the potential problems that can arise from attempting to optimise multiple variables simultaneously, a hierarchical approach to tracking may be adopted whereby proximal variables are optimised initially before tracking distal variables such as limb angles. Given the success of the method presented here for rigid body tracking some confidence can be held for such an extension of the method to articulated motion.

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. Los Alamitos, CA, USA: IEEE Computer Society.

Golland, P. & Bruckstein, A.M. (1997) Motion from color. *Computer Vision and Image Understanding*, **68**, 346-362.

Hel-Or, Y. & Werman, M. (1995) Pose estimation by fusing noisy data of different dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**, 195-201.

Jones, P.R.M. & West, G.M. (1989). Anthropometric and clothing measurement methods. *Technology Review*, **2**, 14.

Khotanzad, A. & Liou, J.J.-H. (1996). Recognition and pose estimation of unoccluded three-dimensional objects from a two-dimensional perspective view by banks of neural networks. *IEEE Transactions on Neural Networks*, **7**, 897-906.

Li, Z. & Wang, H. (1999) Real-time 3D motion tracking with known geometric models. *Real-Time Imaging*, **5**, 167-187.

McKenna, S.J., Raja, Y. & Gong, S. (1999) Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, **17**, 225-231.

Plänkers, R. & Fua, P. (2001) Tracking and modeling people in video sequences. *Computer Vision and Image Understanding*, **81**, 285-302.

Stephens, R.S. (1990) Real-time 3D object tracking. *Image and Vision Computing*, **8**, 91-96.

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.