TOWARDS A MARKER-LESS HUMAN GAIT ANALYSIS SYSTEM

Karl J. Sharman, Mark S. Nixon and John N. Carter
Electronics and Computer Science, University of Southampton, SO17 1BJ, UK
{kjs98r, msn, jnc}@ecs.soton.ac.uk

KEY WORDS: Human Gait Analysis, Dynamic Object Analysis, Volume Intersection, 3D Reconstruction, Computer Vision

ABSTRACT

A non–invasive system is described which is capable of extracting and describing the three–dimensional nature of human gait thereby extending the use of gait as a biometric. Of current three–dimensional systems, those using multiple views appear to be the most suitable. Reformulating the three–dimensional analysis algorithm known as Volume Intersection as an evidence gathering process for moving object extraction gives means to overcome concavities and to handle noise and occlusion. Results on synthetic imagery show that the technique does indeed process a multi-view image sequence to derive the parameters of interest thereby providing a suitable basis for future development as a marker-less gait analysis system.

1 INTRODUCTION

For many years it has been observed that we can identify others by the manner with which they walk; Shakespeare makes several such claims, for example, in *Julius Caesar, ACT I, Scene iii*

> **Casca** Stand close awhile, for here comes one in haste.
>
> **Cassius** 'Tis Cinna; I do know him by his gait;
> He is a friend.

Such claims are now also backed by psychologists whose experiments, commonly using dynamic point–light arrays indicate that individuals can indeed be recognised by their gait. (Nixon et al., 1999) surveys the use of gait as a biometric.

Generalised application for recognising people by their gait mandates research into the development of a three–dimensional analysis system. This system must be able to handle known factors, especially that gait is inherently self occluding: one leg can obscure the other; arms (and apparel) can hide the legs. Further, for recognition to be of application potential, we require a system that is non–invasive, without subject contact. Finally, it is not unlikely that recognition by gait will encounter images of poor quality (as in surveillance videos) suggesting capability to handle noise should be considered at the outset.

The main approaches to three–dimensional vision are well established now. They include laser based range systems, but these are often unsuited to an application which includes the head. Further, application requirements show that video is more suitable, obviating use of laser systems. The video–based three–dimensional systems include shape from a single frame, by texture, defocus, or shading. These all require known assumptions concerning the environment. On the other hand, more recent multiple view techniques appear to be superseding stereo in applications requiring high accuracy with better ability to resolve hidden surfaces. As such, a technique based on multiple views would appear preferential for contactless three–dimensional gait analysis.

Earlier work by Martin and Aggarwal (Martin and Aggarwal, 1983) demonstrated a means of combining multiple calibrated views to generate a surface description of objects. The views were segmented to produce silhouettes whose boundary points were then orthogonally projected through 3D space, producing the *Volume Intersection* (VI). The object was shown to exist within the intersection of these projected views. It was noted that problems existed in the algorithm such as concavities which were later described by the *visual hull* (Laurentini, 1995). Others, including (Potmesil, 1987), used a perspective or conic projection, possibly more appropriate since many of the images used in the investigation of the orthogonal intersection method were actually sourced from perspective views.

Recently a multiview technique has been proposed for three–dimensional moving object analysis. This uses VI separately on each frame and then tracks the object through the sequence (Bottino et al., 1998), but neither uses evidence gathering nor considers the sequence as a whole. We shall show that VI appears to bear close similarity to evidence gathering procedures. This is of especial interest since it is well known that evidence gathering has performance advantages in respect of practical factors discussed earlier, namely an ability to handle noise and occlusion. By performing VI by evidence gathering we will not only confer noise tolerance but also allow ability to accommodate image sequences in their entirety, removing requirement for tracking. We shall first show how VI can be formulated in grey scale (removing the need for segmentation) and as evidence gathering. We will then show how image sequences can be processed to extract mathematically described dynamic three–dimensional objects, with examples being the parameterisation of a moving ball and walking synthetic human in an artificially produced environment.

2 THEORY

2.1 Overview of the approach and system

The approach first requires production of the 3D dynamic data, whose resulting voxel-time-space is analysed to extract and parameterise dynamic objects of interest. Analysis of real world scenes is simplified in the former stage because no assumptions are made about correlation between successive frames of the 2D data. Hence no models are used for the initial analysis, enabling arbitrary scenes to be studied. Models are applied to extract and describe the various objects in the second stage of analysis of the 3D data.

2.2 Gait analysis

Previous work (Cunado et al., 1999) investigated a two–dimensional gait extraction and description method where single lines which oscillated in a pendulum manner were sought. From the edge–detected source images, the pendular lines from the thighs were located by an evidence gathering procedure adapted from the work by (Nash et al., 1997) which itself was an extension to the Hough Transform by (Hough, 1962). No tracking was involved, and the evidence gathering nature of the system made it extremely tolerant to noise and occlusion, with the capability to handle time lapse imagery.

By analysing the frequency components that the pendular lines made with the vertical, as shown in figure 1, a biometric was produced from the phase weighted magnitude spectrum. For the small sample of subjects, this method produced a 100% recognition rate.

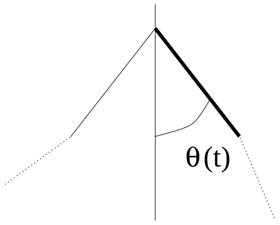


Figure 1: The pendular nature of the thighs used by Cunado et al.

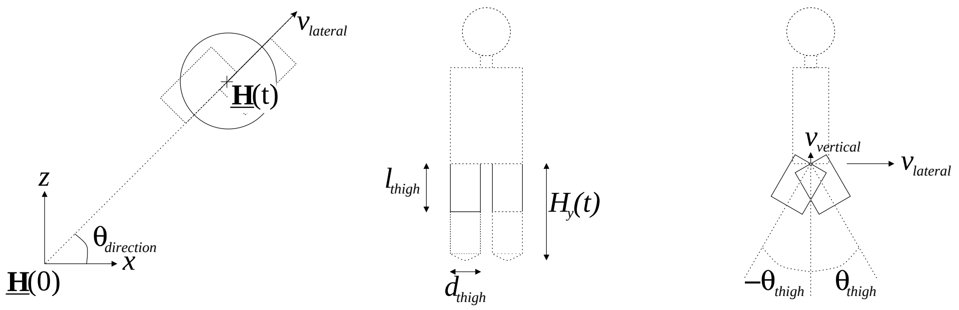
The third–dimension allows for a much more detailed model to be applied to the data, and subjects would also be permitted to walk at any angle to the cameras instead of the simple planar motion used by Cunado, giving true freedom of motion. The swing in the hips and differentiating between the left and right legs are just two examples of the advantage of manipulating the information in 3D. The model that this work uses is currently not capable of distinguishing between the subtleties of unique gait patterns, but by extending the harmonics studied such information will become available for analysis.

2.3 Three dimensional reconstruction

Given an appropriate pin hole camera model, by the principle of collinearity, the relationship between the real and scene coordinates is:

$$\mathbf{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{z}{d} \cdot \begin{bmatrix} x_i \\ y_i \\ d \end{bmatrix} \tag{1}$$

where $(x,\, y,\, z)$ are the real 3D coordinates, $(x_i,\, y_i)$ are the image coordinates, and d is the effective focal length of the camera. In terms of an accumulator of voxels in real space, for all possible values of z we can compute Equation 1. This votes for existence of a voxel in the real space. In VI, only source pixels labelled from a specific feature in the scene are



(a) plan view (b) from the front (c) from the side

Figure 2: The new three–dimensional basic human gait model constructed from spheres and cylinders.

Parameter	Description
H_{x0},H_{y0}, H_{z0}	Hip mean position at time $t = 0$
H_{width}	Hip width
θ_d	Direction of general motion
s	Step rate (steps/second)
H_v	Hip mean speed in the direction of motion
H_{va1}, H_{vb1}	First harmonic pair giving the deviation in the formation motion of the hips
H_{ya1}, H_{yb1}	First harmonic pair giving the deviation in the y position of the hips
T_0	Mean angle of the hips
T_{a1}, T_{b1}	First harmonic pair giving the deviation in the angle of the thighs

Table 1: The 14 parameters required for simple gait recognition.

processed. These pixels are projected through space and their intersection is recovered to describe the object. Concavities cannot be resolved by this, as they lead to a visual hull – volumes within the object that cannot be observed. Evidently VI is constrained to indicate existence only.

To increase descriptive capability, grey scale can be incorporated into a new voting process. The hypothesis is that rays from a point in space will be of a similar level of intensity. For each voxel, its shade, and the confidence in the shade is calculated from the rays which pass through it, based upon the statistical measure:

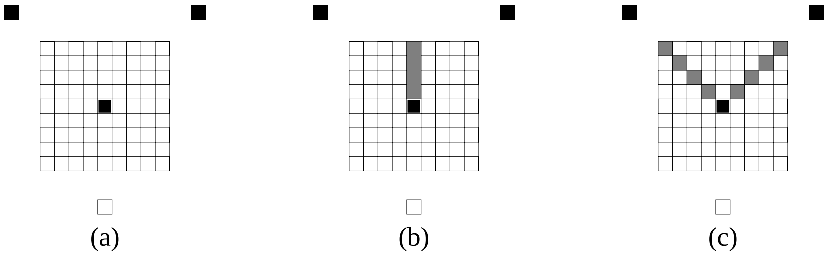
$$m = \frac{\sigma^2 + k_1}{n + k_2} \tag{2}$$

where σ^2 is the variance of the grey level of the n contributing rays. k_1 and k_2 adjust the weight of voting for more views. This measure, based upon the variance, is suited to the reduction of additive noise, and increases as confidence decreases.

The scene is refined by multiple iterations, with each pass selecting remaining voxels in which there is the highest confidence level (a low value of m). The actual shade assigned to the voxel however is not the mean of the n contributing rays, but is selected by averaging the values of pixels which lie closest to the mean, see Figure 3. This is performed so that rays which clearly do not contribute the same information about the voxel do not influence its shade. For such rays, it is then known that another voxel must lie between the respective source views and the selected voxel, but the rays can no longer vote for any voxels beyond the occluding one. For rays which contributed to a voxel's shade, it is assumed that all voxels between the respective source views and the selected voxel are transparent. The source of such rays is then no longer able to produce any further contribution to the reconstruction.

Returning to the hypothesis, it becomes apparent that rays from opposite directions falling onto a voxel can neither vote against each other nor vote with each other. Due to this, the voxels are allocated six sides during the search for those with the highest confidence levels. However, once such voxels are found, the implemented algorithm represents them by only one shade, taken from the side with the greatest certainty, for simplicity of later analysis.

The hypothesis thus dictates suitable camera positions – placing four cameras equispaced around a plane containing the object would not lead to any correlation between views; however if all four cameras were to look down onto the object, then correlation can be made as all views would be able to correlate information regarding the tops of voxels. Since the



(a) 3 sources, producing 2 contributing rays and 1 non-contibuting ray to a voxel
(b) Voxels occluded to the non-contributing source.
(c) Transparent voxels.

Figure 3: Contributing and non–contributing rays.

voxels are of finite size, when they are mapped onto the two–dimensional views they may be represented by more than one pixel. To reduce problems of aliasing, for such voxels, it is from the combined contribution of these pixels that the ray is generated. The rays are recast, with occlusion by previously selected voxels affecting the voting procedure. The votes in each voxel are thus proportional to the statistical match between respective source pixels from the different viewpoints, hence offering facility to resolve concavities. Further, the approach is phrased as evidence gathering, giving requisite performance advantages.

2.4 Three dimensional dynamic object analysis

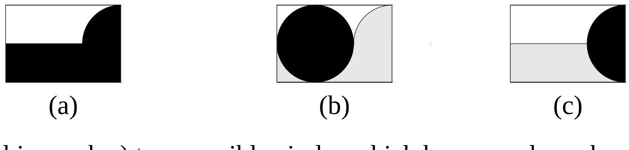
2.4.1 Analysis of a moving sphere In order to accommodate moving objects, we require to integrate multiple views of a sequence of frames. By restricting to moving parametrically described objects, the motion of the centre of a moving sphere is described by:

$$\mathbf{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} s_x \\ s_y \\ s_z \end{bmatrix} + t \cdot \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \tag{3}$$

where $[s_x s_y s_z]^T$ are the sphere's initial coordinates, and $[v_x v_y v_z]^T$ describe velocity along the three axes respectively.

For an accumulator space in terms of velocity and starting position, votes are given to parameter combinations of the moving sphere that could have caused a voxel to be present in the scene, previously recovered in three–dimensions by the above technique. This is performed in a manner similar to the earlier approach (Nash et al., 1997).

However this voting mechanism is unfair for the analysis of small objects or of those which appear in the captured 3D view for only a few of the frames. For example, if a black circle was being sought in Figure 4a, then it would just as likely pick the circle highlighted in Figure 4b, as it would the circle in Figure 4c. The vote therefore needs to be weighted - if the vote was divided by the maximum number of votes that a circle at the position could obtain, then the circle in Figure 4b would receive half the value of the circle in Figure 4c. However this cannot be done if the radius of the circle is unknown; a circle of radius 1 would be found and gain a high vote wherever the image is black. Conversely, without weighting, the best circle would be one with the largest radius.



a) the original image b,c) two possible circles which have equal numbers of votes.

Figure 4: Non–weighted voting for circles.

Two–dimensional tools can be extended for application on the three–dimensional data. Edge detectors which highlight sharp changes in colour or in certainty can be useful, however volume methods are preferable if wider peaks in the accumulator array are required. As with two–dimensional analysis, thresholding the colour obtained is not ideal due to variable lighting conditions in the scene. To aid any recognition, two–dimensional tools such as median filters can be applied to remove background noise, leaving just the moving objects in the scene.

2.4.2 Human gait analysis Increasing the complexity of the mathematical models exponentially increases the size of the accumulator array, thus constructing the accumulator space is infeasible, except for basic models. However genetic algorithms provide a means to search such large dimensional spaces and although their approach is inherently random, with careful use they will converge on the peak of the accumulator, whilst avoiding its construction. This can be aided by ensuring that the peaks will be wide, aided by either blurring the source data, or by searching for volumes rather than edges. The genetic algorithm designed was first used to analyse a moving sphere, and then developed to analyse the walking human model in figure 2. The fourteen parameters for this gait model are described in table 1.

The central position of the hips is thus given by:

$$\mathbf{H} = \begin{bmatrix} H_x \\ H_y \\ H_z \end{bmatrix} = \begin{bmatrix} H_{x0} \\ H_{y0} \\ H_{z0} \end{bmatrix} + \begin{bmatrix} (H_v t + H_{va1}\cos(2st) + H_{vb1}\sin(2st)) \cdot \cos(\theta_d) \\ H_{ya1}\cos(2st) + H_{yb1}\sin(2st) \\ (H_v t + H_{va1}\cos(2st) + H_{vb1}\sin(2st)) \cdot \sin(\theta_d) \end{bmatrix} \tag{4}$$

and the angle of the hips (with the right being the negative angle of the left) given by:

$$\theta_T = T_0 + T_{a1}\cos(st) + T_{b1}\sin(st) \tag{5}$$

Note that the hip harmonics are at twice the frequency of those in the thighs due to the addition of sinusoidal motion caused by the two legs.

The template for the genetic algorithm consists of the two thighs modelled by cylinders, whose length is determined from the maximum height of the hips, and whose radius is determined from the hip width. The thighs are rotated by the angle θ_T about an axis through the hip.

3 RESULTS

3.1 Three dimensional reconstruction results

The three–dimensional reconstruction algorithm, being influenced by the brightness of the pixels from the various views, enables the visual hull problem of the silhouette based VI algorithm to be overcome. An example of this is shown in the open box in Figure 5a. When using a silhouette based approach, no view would be able to highlight the small cube in the corner of the box as it lies completely within the concavity. However, as the cube is of a slightly different intensity, our new reconstruction algorithm can locate the box, as highlighted in Figure 5b.

(a) Two of the six source views.

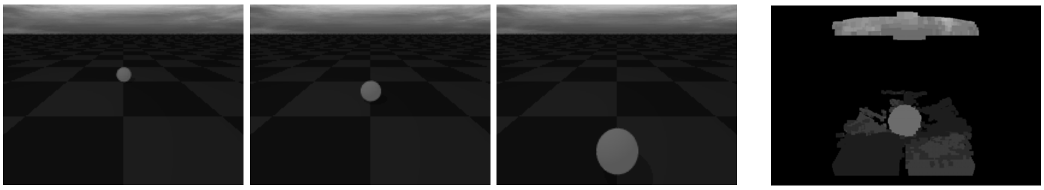
(b) Reconstructed 3D scene.

Figure 5: Open–box source images and resulting scene views.

3.2 Three dimensional dynamic object analysis results

3.2.1 Translating sphere model To illustrate the capture of the parameters of a moving object, a moving sphere was placed in a scene containing a complex sky, and patterned flooring. Ten frames were ray–traced from three views. Figure 6a shows three frames from one of the views. For each triplet of frames, the three–dimensional scene was estimated using our new technique, with the seventh frame presented in Figure 6b.

This set of three–dimensional frames was then analysed using the moving sphere genetic algorithm technique. No pre–processing was performed on the data. The results of the extraction and analysis of the motion can be seen in Table 2, where a ratio of the number of votes to the square root of the number of possible votes was used for the assessment. For all the sets of parameters tested, the peak of the parameter space was correct, although parameters were only tested to a limited degree of accuracy. The first result, analysed to the greatest accuracy of ± 0.5 unit, has particular interest since the sphere is actually only fully present in five of the ten frames.



(a) 3 sphere images from the same camera.

(b) Reconstructed scene.

Figure 6: Sphere test images and resulting reconstruction.

Source ray-tracer values						Estimated values					
Starting position			Velocity			Starting position			Velocity		
s_x, s_y, s_z			v_x, v_y, v_z			s_x, s_y, s_z			v_x, v_y, v_z		
-40	5	0	60	0	0	-40	5	0	60	0	0
20	5	0	-20	0	0	20	5	0	-20	0	0
0	5	0	-20	20	-20	0	5	0	-20	20	-20
0	0	0	0	30	0	0	0	0	0	30	0

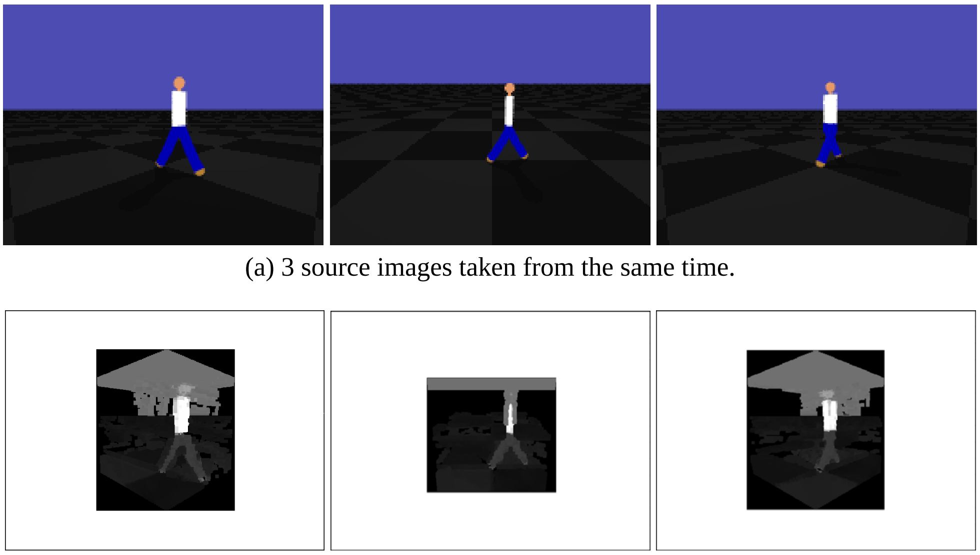
Table 2: Extraction results of a moving sphere of known radius.

3.2.2 Walking human model

Using the 14 dimensional human model above, a synthetic scene was analysed and the gait of the subject extracted. Since the original source parameters are known, this makes an ideal method to test scenes before trials in the real world.

Figure 7 shows a selection of original source images and of the resulting data produced using the three–dimensional analysis algorithm. The anti–aliasing effect can be seen around the reconstructed person in the second view where it coincides with the colour of the background, hence the sweeping of the grey around the person in the first and third views. Increasing the resolution of the voting space would reduce this effect which acts as a large source of error when later estimating the parameters.

The results of a 54 bit genome genetic algorithm successfully reproduce the majority of the original parameters, however a few were inaccurate. Table 3 shows the effect on changing the weighting values. It can be seen that the two different methods are not ideal, and it appears that the exact values lie between those found by the two methods; further mathematical analysis will be performed in this area.



(a) 3 source images taken from the same time.

(b) Reconstructed scene from the same views as in (a)

Figure 7: Synthetic source views and the resulting 3D views of a walking human

Trial	Exact values	Trial using measure $\frac{v}{\sqrt{p}}$	Trial using measure $\frac{v}{p}$
H_0	-12.0, 14.4, 0.0	-12.0, 17.0, 0.0	-12.0, 13.2, 0.0
H_{width}	6.6	7.4	6.2
θ_d	0.0	0.0	0.0
s	1.0	1.0	1.0
H_v	20.0	20.0	20.0
H_{va1} H_{vb1}	0.0, 2.0	0.0, 2.0	-0.5, 2.0
H_{ya1} H_{yb1}	1.6, 0.0	3.6, 1.6	0.2, 1.0
T_0	0.0	0.0	0.0
T_{a1} T_{b1}	0.0, -30.0	0.0, -25.0	0.0, -30.0
Votes,v	3902	5572	3179
Possible votes, p	4714	8326	3654
$\frac{v}{\sqrt{p}}$	58.832	61.065	52.590
$\frac{v}{p}$	0.828	0.669	0.870

Table 3: Extraction results of a walking synthetic human

4 CONCLUSION

We have presented a new formulation of the Volume Intersection method which incorporates grey scale into an evidence gathering technique. It is capable of processing multi–view images of a scene to produce a three–dimensional representation. This algorithm has been shown to be capable of resolving concavities that the silhouette–based VI cannot. The statistical nature of this evidence gathering method now provides noise tolerance to a previously intolerant construction algorithm.

For a synthetic dynamic scene, we have demonstrated that the combination of the three–dimensional scene generation and the parameter estimation algorithms, both of which are evidence gathering techniques, successfully reproduces the parameters for a simple gait model. Increasing the complexity of the model, including incorporating a larger number of harmonics will produce a non–invasive human gait analysis system.

In conclusion, we have therefore described a system that is capable of estimating three–dimensional motion parameters from multi-view images by non–invasive means.

REFERENCES

Bottino, A., Laurentini, A. and Zuccone, P., 1998. Toward non-instrusive motion capture. *Computer Vision – ACCV'98* **2**, pp. 416–423.

Cunado, D., Nash, J., Nixon, M. and Carter, J., 1999. Gait extraction and description by evidence–gathering. *Proceedings of AVBPA 99* pp. 43–48.

Hough, P., 1962. *Method and means for recognising complex patterns*. US Patent, 3,069,654.

Laurentini, A., 1995. How far 3D shapes can be understood from 2D silhouettes. *IEEE Transactions on PAMI* **17**(2), pp. 188–195.

Martin, W. and Aggarwal, J., 1983. Volumetric descriptions of objects from multiple views. *IEEE Transactions on PAMI* **5**(2), pp. 150–158.

Nash, J., Carter, J. and Nixon, M., 1997. *Dynamic feature extraction via the velocity Hough transform*. Pattern Recognition Letters **18**, pp. 1035–1047.

Nixon, M., Carter, J., Cunado, D., Huang, P. and Stevenage, S., 1999. Automatic Gait Recognition. *BIOMETRICS: Personal Identification in Networked Society.* Editors A.K. Jain et al. Kluwer Academic Publishing. pp231–250.

Potmesil, M., 1987. Generating octree models of 3D objects from their silhouettes in a sequence of images. *Computer Vision, Graphics and Image Processing* **40**, pp. 1–29.